Bernoulli **18**(1), 2012, 177–205
DOI: 10.3150/10-BEJ331

Projection-type estimation for varying coefficient regression models

YOUNG K. LEE[1], ENNO MAMMEN[2] and BYEONG U. PARK[3]

[1] *Department of Statistics, Kangwon National University, Chuncheon 200-701, Korea. E-mail: youngklee@kangwon.ac.kr*
[2] *Department of Economics, University of Mannheim, L7, 3-5, 688131 Mannheim, Germany. E-mail: emammen@rumms.uni-mannheim.de*
[3] *Department of Statistics, Seoul National University, Seoul 151-747, Korea. E-mail: bupark@stats.snu.ac.kr*

In this paper we introduce new estimators of the coefficient functions in the varying coefficient regression model. The proposed estimators are obtained by projecting the vector of the full-dimensional kernel-weighted local polynomial estimators of the coefficient functions onto a Hilbert space with a suitable norm. We provide a backfitting algorithm to compute the estimators. We show that the algorithm converges at a geometric rate under weak conditions. We derive the asymptotic distributions of the estimators and show that the estimators have the oracle properties. This is done for the general order of local polynomial fitting and for the estimation of the derivatives of the coefficient functions, as well as the coefficient functions themselves. The estimators turn out to have several theoretical and numerical advantages over the marginal integration estimators studied by Yang, Park, Xue and Härdle [*J. Amer. Statist. Assoc.* **101** (2006) 1212–1227].

Keywords: kernel smoothing; local polynomial regression; marginal integration; oracle properties; smooth backfitting; varying coefficient models

1. Introduction

In this paper we consider a varying coefficient regression model proposed by Hastie and Tibshirani [12] and studied by Yang, Park, Xue and Härdle [24]. The model takes the form $Y^i = m(\mathbf{X}^i, \mathbf{Z}^i) + \sigma(\mathbf{X}^i, \mathbf{Z}^i)\varepsilon^i$, $i = 1, \ldots, n$, where

$$m(\mathbf{X}, \mathbf{Z}) = \sum_{j=1}^{d} m_j(X_j) Z_j, \tag{1.1}$$

m_j are unknown coefficient functions, $\mathbf{X}^i = (X_1^i, \ldots, X_d^i)^\top$ and $\mathbf{Z}^i = (Z_1^i, \ldots, Z_d^i)^\top$ are observed vectors of covariates, and ε^i are the error variables such that $E(\varepsilon^i | \mathbf{X}^i, \mathbf{Z}^i) = 0$ and

1350-7265 © 2012 ISI/BS

$\operatorname{var}(\varepsilon^i|\mathbf{X}^i,\mathbf{Z}^i)=1$. We assume that $(\mathbf{X}^i,\mathbf{Z}^i,Y^i)$ for $1\le i\le n$ are independent and identically distributed. The model is simple in structure and easily interpreted, yet flexible, since the dependence of the response variable on the covariates is modeled in a nonparametric way. The model is different from the functional coefficient model of Chen and Tsay [4], Fan and Zhang [8], Cai, Fan and Li [2] and Cai, Fan and Yao [3], where m_j are functions of a single variable, that is, $m(X^i,\mathbf{Z}^i)=\sum_{j=1}^d m_j(X^i)Z_j^i$. Fitting the latter model is much simpler than the model (1.1) since it involves only a univariate smoothing across the single variable X.

To fit the model (1.1), we may apply the idea of local polynomial smoothing. To illustrate the difficulty in fitting the model, suppose that we employ local constant fitting so that we minimize

$$\sum_{i=1}^n\left[Y^i-\sum_{j=1}^d\theta_j Z_j^i\right]^2 K_h(x_1,X_1^i)\cdots K_h(x_d,X_d^i)$$

with respect to θ_j, $1\le j\le d$, to get estimators of $m_j(x_j)$, $1\le j\le d$, where K_h is a kernel function. For each coefficient m_j, this yields an estimator which is a function of not only x_j but also other variables x_k, $k\ne j$. The marginal integration method, proposed and studied by Yang *et al.* [24], is simply to take the average of $\hat\theta_j(X_1^i,\ldots,X_{j-1}^i,x_j,X_{j+1}^i,\ldots,X_d^i)$ in order to eliminate the dependence on the other variables.

In this paper we propose a new method for fitting the model (1.1). The proposed method is to project the vector of the full-dimensional kernel-weighted local polynomial estimators ($\hat\theta_j$, $1\le j\le d$, in the above, in the case of local constant fitting) onto a space of vectors of functions $f_j:\mathbb{R}\to\mathbb{R}$, $1\le j\le d$, with a suitable norm. Projection-type estimation has been studied in other structured nonparametric regression models. For example, the smooth backfitting method was proposed by Mammen, Linton and Nielsen [17] to fit additive regression models. The same idea was applied to quasi-likelihood additive regression by Yu, Park and Mammen [25] and to additive quantile regression by Lee, Mammen and Park [16]. Some nonparametric time series models have been proposed with unobserved factors Z_j that do not depend on the individual but on time; see, for example, Connor, Linton and Hagmann [5], Fengler, Härdle and Mammen [9] and Park, Mammen, Härdle and Borak [21]. In these papers it has been shown that one can also proceed asymptotically in the models under consideration, as if the factors would have been observed. We note that the current problem does not fit into the framework of the above papers but requires a different treatment. In particular, in the model (1.1), the functions m_j are not additive components of the regression function, but they are the coefficients of Z_j. For a treatment of our model we have to exclude the case of constant $Z_j\equiv 1$. In the case of constant Z_j, model (1.1) reduces to the additive model. The key element in the derivation of the theory for our model is to embed the vector of the coefficient functions into an additive space of vectors of univariate functions and then to endow the space with a norm where the covariates Z_j enter with kernel weights.

As far as we know, the marginal integration method has been the only method to fit the model (1.1). It is widely accepted that the marginal integration method suffers from the curse of dimensionality. Inspired by Fan, Härdle and Mammen [6] and others, Yang *et al.* [24] tried to avoid the dimensionality problem by using two different types of kernels and bandwidths. To be more specific, consider estimation of m_j for a particular j. The method then uses a kernel, say L, and bandwidths, say b_k, for the directions of x_k $(k \neq j)$, which are different from a kernel K and a bandwidth h_j for the direction of x_j. By choosing $b_k \ll h_j$ and taking a higher order kernel L, we can achieve the univariate optimal rate of convergence for the resulting estimator of m_j. One of the main difficulties with the marginal integration method is that there is no formula available for the optimal choice of the secondary bandwidths b_k. Also, the performance of the method depends crucially on the choice of the secondary bandwidths b_k, as observed in our numerical study; see Section 5. Furthermore, the method involves estimation of a full-dimensional regression estimator, which requires inversion of a full-dimensional $[(\pi+1)d] \times [(\pi+1)d]$ smoothing matrix, where π is the order of local polynomial fitting. This means that the method may break down in practice in high dimension.

On the contrary, the proposed method may use bandwidths of the same order for all directions to achieve the univariate optimal rate of convergence, and we derive formulas for the optimal bandwidths. The method requires only one- and two-dimensional smoothing and inversion of a $(\pi+1) \times (\pi+1)$ matrix which is computed by a single-dimensional local smoothing. Thus, the proposed method does not suffer from the curse of dimensionality in practice as well as in theory. We show that the method has the oracle properties, meaning that the proposed estimator of m_j for each j has the same first-order asymptotic properties as the oracle (infeasible) estimator of m_j that uses the knowledge of all other coefficient functions m_k, $k \neq j$. We develop the theory for the method with local polynomial fitting of general order $\pi \geq 0$. Thus, the theory gives the asymptotic distributions of the estimators of m_j, as well as their derivatives $m_j^{(k)}$, $1 \leq k \leq \pi$.

There have been several works on a related varying coefficient model where the coefficients are time-varying functions. These include Hoover, Rice, Wu and Yang [13], Huang, Wu and Zhou [14, 15], Wang, Li and Huang [23] and Noh and Park [19]. The kernel method of fitting this model is quite different from, and simpler than, the method of fitting our model (1.1), since the former involves only a univariate smoothing across time. Recently, Park, Hwang and Park [20] considered a testing problem for the model (1.1) based on the marginal integration method.

This paper is organized as follows. In the next section, we describe the proposed method with local constant fitting and then, in Section 3, we give its theoretical properties. Section 4 is devoted to the extension of the method and theory to local polynomial fitting of general order. In Section 5 we present the results of our numerical study. In Section 6 we apply the proposed method to Boston Housing Data. Technical details are contained in the Appendix.

2. The method with local constant fitting

Although our main focus is to introduce the method with local polynomial fitting and to develop its general theory, we start with local constant fitting since the method is better understood in the latter setting. Let Y be the response variable, and $\mathbf{X} = (X_1, \ldots, X_d)^\top$

and $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top$ be the covariate vectors of dimension d. Let $\{(\mathbf{X}^i, \mathbf{Z}^i, Y^i)\}_{i=1}^n$ be a random sample drawn from $(\mathbf{X}, \mathbf{Z}, Y)$. Assume that the density p of $\mathbf{X}$ is supported on $[0,1]^d$. To estimate the coefficient functions m_j in the model (1.1), we consider a 'smoothed' squared error loss. Similar ideas were adopted for additive regression by Mammen *et al.* [17] and for quasi-likelihood additive regression by Yu *et al.* [25].

Let K be a nonnegative function, called a *base kernel*. To define a smoothed squared error loss, we use a boundary corrected kernel, as in Mammen *et al.* [17] and Yu *et al.* [25], which is defined by

$$K_g(u,v) = \left[\int_0^1 K\left(\frac{w-v}{g}\right) \mathrm{d}w\right]^{-1} K\left(\frac{u-v}{g}\right) I(u, v \in [0,1]).$$

Suppose that we use different bandwidths for different directions. Let $\mathbf{h} = (h_1, \ldots, h_d)$ be the bandwidth vector. For simplicity, we focus on the case where we use a product kernel of the form $K_{\mathbf{h}}(\mathbf{u}, \mathbf{v}) = \prod_{j=1}^d K_{h_j}(u_j, v_j)$. We may use a more general multivariate kernel, but this would require more involved notation and technical arguments. The proposed estimator of $\mathbf{m} \equiv (m_1, \ldots, m_d)^\top : \mathbb{R}^d \to \mathbb{R}^d$, where $m_j(\mathbf{x}) = m_j(x_j)$, is defined to be the minimizer of

$$L(\mathbf{f}) = \int n^{-1} \sum_{i=1}^n \left[Y^i - \sum_{j=1}^d f_j(x_j) Z_j^i\right]^2 K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i)\, \mathrm{d}\mathbf{x}$$

over $\mathbf{f} = (f_1, \ldots, f_d)^\top$ with $L(\mathbf{f}) < \infty$. Here and hereafter, integration over $\mathbf{x}$ is on $[0,1]^d$. Define $\hat{\mathbf{M}}(\mathbf{x}) = n^{-1} \sum_{i=1}^n K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i) \mathbf{Z}^i \mathbf{Z}^{i\top}$. Then, $L(\mathbf{f}) < \infty$ is equivalent to $\int \mathbf{f}(\mathbf{x})^\top \hat{\mathbf{M}}(\mathbf{x}) \mathbf{f}(\mathbf{x})\, \mathrm{d}\mathbf{x} < \infty$. The function space that arises in the minimization problem is

$$\mathcal{H}(\hat{\mathbf{M}}) = \{\mathbf{f} \in L_2(\hat{\mathbf{M}}) : f_j(\mathbf{x}) = g_j(x_j) \text{ for a function } g_j : \mathbb{R} \to \mathbb{R}, 1 \le j \le d\},$$

where $L_2(\hat{\mathbf{M}})$ denotes a class of function vectors $\mathbf{f}$ defined by

$$L_2(\hat{\mathbf{M}}) = \left\{\mathbf{f} : \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}))^\top \text{ for some functions } f_j : \mathbb{R}^d \to \mathbb{R} \right.$$
$$\left. \text{and } \int \mathbf{f}(\mathbf{x})^\top \hat{\mathbf{M}}(\mathbf{x}) \mathbf{f}(\mathbf{x})\, \mathrm{d}\mathbf{x} < \infty \right\}.$$

The spaces $L_2(\hat{\mathbf{M}})$ and $\mathcal{H}(\hat{\mathbf{M}})$ are Hilbert spaces equipped with a (semi)norm $\|\cdot\|_{\hat{\mathbf{M}}}$, defined by

$$\|\mathbf{f}\|_{\hat{\mathbf{M}}}^2 = \int \mathbf{f}(\mathbf{x})^\top \hat{\mathbf{M}}(\mathbf{x}) \mathbf{f}(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Let $\mathbf{M}(\mathbf{x}) = E(\mathbf{Z}\mathbf{Z}^\top | \mathbf{X} = \mathbf{x}) p(\mathbf{x})$. Since $\|\mathbf{f}\|_{\hat{\mathbf{M}}}^2$ converges to

$$\|\mathbf{f}\|_{\mathbf{M}}^2 \equiv \int \mathbf{f}(\mathbf{x})^\top \mathbf{M}(\mathbf{x}) \mathbf{f}(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

in probability under certain conditions, the corresponding Hilbert spaces in the limit are $L_2(\mathbf{M})$ and $\mathcal{H}(\mathbf{M})$, which are defined as $L_2(\hat{\mathbf{M}})$ and $\mathcal{H}(\hat{\mathbf{M}})$, respectively, with $\hat{\mathbf{M}}$ being replaced by $\mathbf{M}$. Here, we note that $\|\cdot\|_{\mathbf{M}}$ becomes a norm if we assume that

$$\mathbf{f}(\mathbf{X})^\top \mathbf{Z} = 0 \text{ almost surely implies } \mathbf{f} = \mathbf{0}. \tag{2.1}$$

In fact, the assumption (2.1) is known to be a sufficient condition for avoiding *concurvity*, as termed by Hastie and Tibshirani [11], an analog of collinearity in linear models. If the assumption does not hold, then the m_j are not identifiable. This is because, for $\mathbf{f}$ such that $\mathbf{f}(\mathbf{X})^\top \mathbf{Z} = 0$ almost surely, we have

$$E(Y|\mathbf{X},\mathbf{Z}) = \mathbf{m}(\mathbf{X})^\top \mathbf{Z} = [\mathbf{m}(\mathbf{X}) + \mathbf{f}(\mathbf{X})]^\top \mathbf{Z}.$$

The assumption (2.1) is satisfied if we assume that the smallest eigenvalue of $E(\mathbf{Z}\mathbf{Z}^\top | \mathbf{X} = \mathbf{x})$ is bounded away from zero on $[0,1]^d$.

For $\mathbf{f} \in \mathcal{H}(\hat{\mathbf{M}})$, we obtain

$$\begin{aligned} L(\mathbf{f}) = &\int n^{-1} \sum_{i=1}^n [Y^i - \tilde{\mathbf{m}}(\mathbf{x})^\top \mathbf{Z}^i]^2 K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i)\, d\mathbf{x} \\ &+ \int [\tilde{\mathbf{m}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})]^\top \hat{\mathbf{M}}(\mathbf{x}) [\tilde{\mathbf{m}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})]\, d\mathbf{x}, \end{aligned}$$

where $\tilde{\mathbf{m}}$ is the minimizer of $L(\mathbf{f})$ over $\mathbf{f} \in L_2(\hat{\mathbf{M}})$. It is given explicitly as

$$\tilde{\mathbf{m}}(\mathbf{x}) = \hat{\mathbf{M}}(\mathbf{x})^{-1} n^{-1} \sum_{i=1}^n \mathbf{Z}^i Y^i K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i). \tag{2.2}$$

Thus, the proposed estimator $\hat{\mathbf{m}} = (\hat{m}_1, \dots, \hat{m}_d)^\top$ can be defined equivalently as the projection of $\tilde{\mathbf{m}}$ onto $\mathcal{H}(\hat{\mathbf{M}})$:

$$\hat{\mathbf{m}} = \operatorname*{argmin}_{\mathbf{f} \in \mathcal{H}(\hat{\mathbf{M}})} \|\tilde{\mathbf{m}} - \mathbf{f}\|_{\hat{\mathbf{M}}}^2. \tag{2.3}$$

By considering the Gâteaux or Fréchet derivatives of the objective function with respect to $\mathbf{f}$, the solution $\hat{\mathbf{m}}$ of the minimization problem (2.3) satisfies the following system of integral equations:

$$0 = \int \hat{\mathbf{M}}_j(\mathbf{x})^\top [\tilde{\mathbf{m}}(\mathbf{x}) - \hat{\mathbf{m}}(\mathbf{x})]\, d\mathbf{x}_{-j}, \qquad 1 \le j \le d, \tag{2.4}$$

where $\hat{\mathbf{M}}_j$ are defined by $\hat{\mathbf{M}} = (\hat{\mathbf{M}}_1, \dots, \hat{\mathbf{M}}_d)^\top$ and $\mathbf{x}_{-j} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)^\top$. In fact, the system (2.4) turns out to be a backfitting system of equations. To see this, we define

$$\tilde{m}_j(x_j) = \hat{q}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_j^i) Z_j^i Y^i,$$

$$\begin{aligned}
\hat{q}_j(x_j) &= n^{-1}\sum_{i=1}^{n} K_{h_j}(x_j, X_j^i)(Z_j^i)^2,\\
\hat{q}_{jk}(x_j, x_k) &= n^{-1}\sum_{i=1}^{n} K_{h_j}(x_j, X_j^i)K_{h_k}(x_k, X_k^i)Z_j^i Z_k^i, \qquad k \neq j.
\end{aligned}$$

We note that, by definition, $\tilde{m}_j : \mathbb{R} \to \mathbb{R}$ does not equal the jth component of $\tilde{\mathbf{m}}$, which maps $\mathbb{R}^d$ to $\mathbb{R}$. We can then see that

$$\begin{aligned}
&\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \tilde{\mathbf{m}}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} = \tilde{m}_j(x_j)\hat{q}_j(x_j),\\
&\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \hat{\mathbf{m}}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} = \hat{m}_j(x_j)\hat{q}_j(x_j) + \sum_{k=1,\neq j}^{d} \int \hat{m}_k(x_k)\hat{q}_{jk}(x_j, x_k)\,\mathrm{d}x_k.
\end{aligned}$$

The second formula is obtained by using the following property of the boundary corrected kernel: $\int K_{h_j}(u_j, v_j)\,\mathrm{d}u_j = 1$. Thus, the system of equations (2.4) is equivalent to

$$\hat{m}_j(x_j) = \tilde{m}_j(x_j) - \sum_{k=1,\neq j}^{d} \int \hat{m}_k(x_k)\frac{\hat{q}_{jk}(x_j, x_k)}{\hat{q}_j(x_j)}\,\mathrm{d}x_k, \qquad 1 \le j \le d. \tag{2.5}$$

We emphasize that the proposed method does not require computation of the full-dimensional estimator $\tilde{\mathbf{m}}(\mathbf{x})$ at (2.2). It only requires one- and two-dimensional smoothing to compute $\tilde{m}_j$, $\hat{q}_j$ and $\hat{q}_{jk}$, and involves inversion of $\hat{q}_j$ only. In contrast, the marginal integration method studied by Yang *et al.* [24] involves the computation of $\tilde{\mathbf{m}}(\mathbf{x})$, which requires inversion of the full-dimensional smoothing matrix $\hat{\mathbf{M}}$. Thus, in practice, the marginal integration method may break down in high dimensions where d is large.

We express the updating equations (2.5) in terms of projections onto suitable function spaces. This representation is particularly useful in our theoretical development. We consider $\mathcal{H}_j(\hat{\mathbf{M}})$, $1 \le j \le d$, subspaces of $\mathcal{H}(\hat{\mathbf{M}})$ defined by

$$\mathcal{H}_j(\hat{\mathbf{M}}) = \{\mathbf{f} \in L_2(\hat{\mathbf{M}}) : f_j(\mathbf{x}) = g_j(x_j) \text{ for a function } g_j : \mathbb{R} \to \mathbb{R}, f_k \equiv 0 \text{ for } k \neq j\}.$$

With this definition, we have

$$\mathcal{H}(\hat{\mathbf{M}}) = \mathcal{H}_1(\hat{\mathbf{M}}) + \cdots + \mathcal{H}_d(\hat{\mathbf{M}}).$$

Also, denoting the projection operator onto a closed subspace $\mathcal{H}$ of $\mathcal{H}(\hat{\mathbf{M}})$ by $\Pi(\cdot|\mathcal{H})$ and its jth element by $\Pi(\cdot|\mathcal{H})_j$, we get, for $\mathbf{f} \in L_2(\hat{\mathbf{M}})$,

$$\begin{aligned}
&\Pi(\mathbf{f}|\mathcal{H}_j(\hat{\mathbf{M}}))_j = \hat{q}_j(x_j)^{-1}\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j},\\
&\Pi(\mathbf{f}|\mathcal{H}_j(\hat{\mathbf{M}}))_k = 0, \qquad k \neq j.
\end{aligned} \tag{2.6}$$

In particular, for $\mathbf{f} \in \mathcal{H}(\hat{\mathbf{M}})$, we have

$$\Pi(\mathbf{f}|\mathcal{H}_j(\hat{\mathbf{M}}))_j = f_j(x_j) + \sum_{k=1,\neq j}^{d} \int f_k(x_k) \frac{\hat{q}_{jk}(x_j, x_k)}{\hat{q}_j(x_j)} \, \mathrm{d}x_k. \tag{2.7}$$

Furthermore, for $\mathbf{f} \in \mathcal{H}_k(\hat{\mathbf{M}})$,

$$\Pi(\mathbf{f}|\mathcal{H}_j(\hat{\mathbf{M}}))_j = \int f_k(x_k) \frac{\hat{q}_{jk}(x_j, x_k)}{\hat{q}_j(x_j)} \, \mathrm{d}x_k, \qquad j \neq k. \tag{2.8}$$

For $\hat{\mathbf{m}} \in \mathcal{H}(\hat{\mathbf{M}})$, let $\hat{\mathbf{m}}_j(\mathbf{x}) = (0, \ldots, 0, \hat{m}_j(x_j), 0, \ldots, 0)^\top$ denote the vector whose jth entry equals $\hat{m}_j(x_j)$, the rest being zero. We can then decompose $\hat{\mathbf{m}}$ as $\hat{\mathbf{m}} = \hat{\mathbf{m}}_1 + \cdots + \hat{\mathbf{m}}_d$. From (2.5) and (2.8), we obtain

$$\hat{\mathbf{m}}_j = \Pi\left(\tilde{\mathbf{m}} - \sum_{k=1,\neq j}^{d} \hat{\mathbf{m}}_k \middle| \mathcal{H}_j(\hat{\mathbf{M}})\right), \qquad 1 \leq j \leq d. \tag{2.9}$$

The backfitting equations (2.5), or their equivalent forms (2.9), give the following backfitting algorithm.

Backfitting algorithm. *With a set of initial estimates $\hat{m}_j^{[0]}$, iterate for $r = 1, 2, \ldots$ the following process: for $1 \leq j \leq d$,*

$$\begin{aligned}\hat{m}_j^{[r]}(x_j) = \tilde{m}_j(x_j) &- \sum_{k=1}^{j-1} \int \hat{m}_k^{[r]}(x_k) \frac{\hat{q}_{jk}(x_j, x_k)}{\hat{q}_j(x_j)} \, \mathrm{d}x_k \\ &- \sum_{k=j+1}^{d} \int \hat{m}_k^{[r-1]}(x_k) \frac{\hat{q}_{jk}(x_j, x_k)}{\hat{q}_j(x_j)} \, \mathrm{d}x_k\end{aligned}$$

or, equivalently,

$$\hat{\mathbf{m}}_j^{[r]} = \Pi\left(\tilde{\mathbf{m}} - \sum_{k=1}^{j-1} \hat{\mathbf{m}}_k^{[r]} - \sum_{k=j+1}^{d} \hat{\mathbf{m}}_k^{[r-1]} \middle| \mathcal{H}_j(\hat{\mathbf{M}})\right). \tag{2.10}$$

3. Theoretical properties of the local constant method

3.1. Convergence of the backfitting algorithm

The theoretical development for the backfitting algorithm (2.10) and for the solution of the backfitting equation (2.9) does not fit into the framework of an additive regression function as in Mammen *et al.* [17]. Formally, we get their model by taking $Z_j \equiv 1$ for all $1 \le j \le d$ in (1.1). However, for identifiability of m_j, we need the assumption that $E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})$ is invertible; see the assumption (A1) below. Trivially, this assumption does not hold for the additive model with $Z_j \equiv 1$. For our model, we directly derive the theoretical properties of the algorithm and the estimators by borrowing some relevant theory on projection operators.

Let $\hat{\Pi}_j$ denote the projection operator $\Pi(\cdot|\mathcal{H}_j(\hat{\mathbf{M}}))$ and Π_j the projection operator $\Pi(\cdot|\mathcal{H}_j(\mathbf{M}))$. Define $\hat{Q}_j = I - \hat{\Pi}_j$ and $Q_j = I - \Pi_j$; these are the projection operators onto $\mathcal{H}_j(\hat{\mathbf{M}}))^\perp$ and $\mathcal{H}_j(\mathbf{M}))^\perp$, respectively. From the backfitting algorithm (2.10), it follows that

$$\begin{aligned}\tilde{\mathbf{m}} - \sum_{k=1}^{j} \hat{\mathbf{m}}_k^{[r]} - \sum_{k=j+1}^{d} \hat{\mathbf{m}}_k^{[r-1]} &= \hat{Q}_j \left(\tilde{\mathbf{m}} - \sum_{k=1}^{j-1} \hat{\mathbf{m}}_k^{[r]} - \sum_{k=j+1}^{d} \hat{\mathbf{m}}_k^{[r-1]} \right) \\ &= \hat{Q}_j \left(\tilde{\mathbf{m}} - \sum_{k=1}^{j-1} \hat{\mathbf{m}}_k^{[r]} - \sum_{k=j}^{d} \hat{\mathbf{m}}_k^{[r-1]} \right).\end{aligned} \tag{3.1}$$

Define $\hat{Q} = \hat{Q}_d \cdots \hat{Q}_1$. Repeated application of (3.1) for $j = d, d-1, \ldots, 1$ gives

$$\tilde{\mathbf{m}} - \hat{\mathbf{m}}^{[r]} = \hat{Q}(\tilde{\mathbf{m}} - \hat{\mathbf{m}}^{[r-1]}).$$

This establishes that

$$\hat{\mathbf{m}}^{[r]} = \hat{Q}\hat{\mathbf{m}}^{[r-1]} + \hat{\mathbf{r}} = \sum_{s=0}^{r-1} \hat{Q}^s \hat{\mathbf{r}} + \hat{Q}^r \hat{\mathbf{m}}^{[0]}, \tag{3.2}$$

where $\hat{\mathbf{r}} = (I - \hat{Q})\tilde{\mathbf{m}}$. If we write $\tilde{\mathbf{m}}_j(\mathbf{x}) = (0, \ldots, 0, \tilde{m}_j(x_j), 0, \ldots, 0)^\top$, then $\hat{\Pi}_j \tilde{\mathbf{m}} = \tilde{\mathbf{m}}_j$ so that

$$\begin{aligned}\hat{\mathbf{r}} &= (\hat{\Pi}_d + \hat{Q}_d \hat{\Pi}_{d-1} + \cdots + \hat{Q}_d \cdots \hat{Q}_2 \hat{\Pi}_1)\tilde{\mathbf{m}} \\ &= \tilde{\mathbf{m}}_d + \hat{Q}_d \tilde{\mathbf{m}}_{d-1} + \cdots + \hat{Q}_d \cdots \hat{Q}_2 \tilde{\mathbf{m}}_1.\end{aligned} \tag{3.3}$$

Convergence of the backfitting algorithm (2.10) depends on the statistical properties of the operator $\hat{Q}$. Consider the event $\mathcal{E}_n$, where $\hat{\mathbf{r}}, \hat{\mathbf{m}}^{[0]} \in \mathcal{H}(\mathbf{M})$ and the norm of the operator $\hat{Q}$ is strictly less than one, that is, $\|\hat{Q}\| < 1$. Here and below, for an operator $F : \mathcal{H}(\mathbf{M}) \to \mathcal{H}(\mathbf{M})$,

$$\|F\| = \sup\{\|F\mathbf{f}\|_{\mathbf{M}} : \mathbf{f} \in \mathcal{H}(\mathbf{M}), \|\mathbf{f}\|_{\mathbf{M}} \le 1\}.$$

Then, in that event, $\sum_{s=0}^{\infty} \hat{Q}^s \hat{\mathbf{r}}$ is well defined in $\mathcal{H}(\mathbf{M})$ and, by (3.2), $\hat{\mathbf{m}}^{[r]}$ converges to $\sum_{s=0}^{\infty} \hat{Q}^s \hat{\mathbf{r}}$ as r tends to infinity. The limit is a solution of the backfitting equation (2.9) since the latter is equivalent to $\hat{\mathbf{m}} = \hat{Q}\hat{\mathbf{m}} + \hat{\mathbf{r}}$. Furthermore, the solution is unique since repeated application of $\hat{\mathbf{m}} = \hat{Q}\hat{\mathbf{m}} + \hat{\mathbf{r}}$ leads to $\hat{\mathbf{m}} = \sum_{s=0}^{\infty} \hat{Q}^s \hat{\mathbf{r}}$.

Below, we collect the assumptions that make the event $\mathcal{E}_n$ occur with probability tending to one and state a theorem for the convergence of the backfitting algorithm (2.10).

Assumptions.

(A1) $E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})$ *is continuous and its smallest eigenvalue is bounded away from zero on* $[0,1]^d$.

(A2) $\sup_{\mathbf{x}\in[0,1]^d} E(Z_j^4|\mathbf{X}=\mathbf{x})<\infty$ *for all* $1\le j\le d$.

(A3) *The joint density* p *of* $\mathbf{X}$ *is bounded away from zero and is continuous on* $[0,1]^d$.

(A4) $E|Y|^\alpha<\infty$ *for some* $\alpha>5/2$.

(A5) K *is a bounded and symmetric probability density function supported on* $[-1,1]$ *and is Lipschitz continuous. The bandwidths* h_j *converge to zero and* $nh_j/(\log n)\to\infty$ *as* $n\to\infty$.

The assumption (A1) implies the concurvity condition (2.1) since it implies that there exists a constant $c>0$ such that for $\mathbf{f}\in\mathcal{H}(\mathbf{M})$,

$$\|\mathbf{f}\|^2_{\mathbf{M}}\ge c\sum_{j=1}^d\int f_j(x_j)^2p_j(x_j)\,\mathrm{d}x_j, \tag{3.4}$$

where p_j denotes the marginal density function of X_j. The inequality (3.4) also tells us that the convergence of $\hat{\mathbf{m}}$ in $\mathcal{H}(\mathbf{M})$ implies the convergence of each component m_j in the usual L_2 norm.

Theorem 1. *Assume that* (A1)–(A5) *hold. Then, with probability tending to one, there exists a solution* $\{\hat m_j\}_{j=1}^d$ *of the backfitting equation (2.5) or (2.9) that is unique. Furthermore, there exist constants* $0<\gamma<1$ *and* $0<C<\infty$ *such that, with probability tending to one,*

$$\sum_{j=1}^d\int[\hat m_j^{[r]}(x_j)-\hat m_j(x_j)]^2p_j(x_j)\,\mathrm{d}x_j\le C\gamma^{2r}\sum_{j=1}^d\int[\tilde m_j(x_j)^2+\hat m_j^{[0]}(x_j)^2]p_j(x_j)\,\mathrm{d}x_j.$$

3.2. Asymptotic distribution of the backfitting estimators

Next, we present the asymptotic distributions of $\hat m_j$. Define

$$\tilde{\mathbf{m}}^A(\mathbf{x})=\hat{\mathbf{M}}(\mathbf{x})^{-1}n^{-1}\sum_{i=1}^n\mathbf{Z}^i[Y^i-m(\mathbf{X}^i,\mathbf{Z}^i)]K_{\mathbf{h}}(\mathbf{x},\mathbf{X}^i),$$

where $m(\mathbf{X},\mathbf{Z})$ is as given in (1.1), and let $\tilde{\mathbf{m}}^B=\tilde{\mathbf{m}}-\tilde{\mathbf{m}}^A$. As in the proof of Theorem 1, we can prove that, for $s=A$ or B, there exists a unique solution $\hat{\mathbf{m}}^s\in\mathcal{H}(\hat{\mathbf{M}})$ of the corresponding backfitting equation (2.9) where $\tilde{\mathbf{m}}$ is replaced by $\tilde{\mathbf{m}}^s$. By the uniqueness of $\hat{\mathbf{m}}$, it follows that $\hat{\mathbf{m}}=\hat{\mathbf{m}}^A+\hat{\mathbf{m}}^B$.

Put $\hat{\mathbf{m}}^A=(\hat m_1^A,\dots,\hat m_d^A)^\top\in\mathcal{H}(\hat{\mathbf{M}})$. In the proof of the following theorem, we will show that $\hat m_j^A$ are well approximated by $\tilde m_j^A\equiv(\hat\Pi_j\tilde{\mathbf{m}}^A)_j$. Note that

$$(\hat\Pi_j\tilde{\mathbf{m}}^A)_j(x_j)=\hat q_j(x_j)^{-1}n^{-1}\sum_{i=1}^nZ_j^i[Y^i-m(\mathbf{X}^i,\mathbf{Z}^i)]K_{h_j}(x_j,X_j^i).$$

Assume that the bandwidths h_j are asymptotic to $c_j n^{-1/5}$ for some constants $0 < c_j < \infty$. By the standard techniques of kernel smoothing, it can be proven that, for $\mathbf{x}$ in $(0,1)^d$, $(\tilde m_1^A(x_j), \ldots, \tilde m_d^A(x_d))^\top$, and thus $\hat{\mathbf{m}}^A$, is asymptotically normal with mean zero and variance $n^{-4/5}\operatorname{diag}(v_j(x_j))$, where

$$v_j(x_j) = \frac{E[Z_j^2 \sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j]}{c_j p_j(x_j)[E(Z_j^2|X_j = x_j)]^2} \int K(u)^2 \,\mathrm{d}u$$

and $\sigma^2(\mathbf{X},\mathbf{Z}) = \operatorname{var}(Y|\mathbf{X},\mathbf{Z})$. Here, it is worth noting that the vector $\tilde{\mathbf{m}}^A$, which belongs to $L_2(\hat{\mathbf{M}})$, does not equal $(\tilde m_1^A(x_j), \ldots, \tilde m_d^A(x_d))^\top \in \mathcal{H}(\hat{\mathbf{M}})$.

The bias of the estimator $\hat{\mathbf{m}}$ comes from $\hat{\mathbf{m}}^B$, which is the projection of $\tilde{\mathbf{m}}^B = (\tilde m_1^B, \ldots, \tilde m_d^B)^\top$ onto $\mathcal{H}(\hat{\mathbf{M}})$. Define $\boldsymbol{\eta}(\mathbf{x}) = (c_1^2 m_1''(x_1), \ldots, c_d^2 m_d''(x_d))^\top$ and $\boldsymbol{\beta}_0(\mathbf{x})$ by

$$\boldsymbol{\beta}_0(\mathbf{x}) = \left[\sum_{k=1}^d c_k^2 m_k'(x_k) p(\mathbf{x})^{-1} E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})^{-1} \frac{\partial}{\partial x_k}(E(\mathbf{Z}Z_k|\mathbf{X}=\mathbf{x})p(\mathbf{x})) + \frac{1}{2}\boldsymbol{\eta}(\mathbf{x})\right] \times \int u^2 K(u)\,\mathrm{d}u.$$

Note that $\tilde{\mathbf{m}}$ and $\boldsymbol{\beta}_0$ do not belong to $\mathcal{H}(\mathbf{M})$. In the proof of the next theorem, we will show that $\boldsymbol{\beta}_0(\mathbf{x})$ is the asymptotic bias of $\tilde{\mathbf{m}}(\mathbf{x})$ as an estimator of $\mathbf{m}(\mathbf{x})$ and that the asymptotic bias of $\hat{\mathbf{m}}(\mathbf{x})$ equals $\boldsymbol{\beta}(\mathbf{x})$, where $\boldsymbol{\beta}$ is the projection of $\boldsymbol{\beta}_0$ onto $\mathcal{H}(\mathbf{M})$:

$$\boldsymbol{\beta} \equiv \Pi(\boldsymbol{\beta}_0|\mathcal{H}(\mathbf{M})) = \operatorname*{argmin}_{\mathbf{f}\in\mathcal{H}(\mathbf{M})} \int [\boldsymbol{\beta}_0(\mathbf{x}) - \mathbf{f}(\mathbf{x})]^\top \mathbf{M}(\mathbf{x})[\boldsymbol{\beta}_0(\mathbf{x}) - \mathbf{f}(\mathbf{x})]\,\mathrm{d}\mathbf{x}.$$

We write $\boldsymbol{\beta}(\mathbf{x}) = (\beta_1(x_1), \ldots, \beta_d(x_d))^\top$.

The following theorem, which demonstrates the asymptotic joint distribution of $\hat m_j$, requires an additional condition on m_j.

(A6) $E(\mathbf{Z}\mathbf{Z}^\top\sigma^2(\mathbf{X},\mathbf{Z})|\mathbf{X}=\mathbf{x})$ is continuous on $[0,1]^d$.

(A7) The coefficient functions m_j are twice continuously differentiable on $[0,1]$, and $E(Z_j Z_k|\mathbf{X}=\mathbf{x})$ is continuously partially differentiable on $[0,1]^d$ for all $1 \le j,k \le d$.

Theorem 2. *Assume that* (A1)–(A7) *hold and that the bandwidths h_j are asymptotic to $c_j n^{-1/5}$ for some constants $0 < c_j < \infty$. Then, for any $\mathbf{x} \in (0,1)^d$, $n^{2/5}[\hat m_j(x_j) - m_j(x_j)]$ for $1 \le j \le d$ are jointly asymptotically normal with mean $(\beta_1(x_1), \ldots, \beta_d(x_d))^\top$ and variance* $\operatorname{diag}(v_j(x_j))$.

4. The method with local polynomial fitting

The method we studied in the previous two sections is based on local constant fitting, where we approximate $f_j(X_j^i)$ by $f_j(x_j)$ when X_j^i are near x_j, in the least-squares criterion $\sum_{i=1}^n [Y^i - \sum_{j=1}^d f_j(X_j^i) Z_j^i]^2$. The method may be extended to local polynomial fitting, where we approximate $f_j(X_j^i)$ by $f_j(x_j) + (X_j^i - x_j)f_j^{(1)}(x_j) + \cdots + (X_j^i - x_j)^\pi f_j^{(\pi)}(x_j)/\pi!$ for X_j^i near x_j. Here and below, $g^{(k)}$ denotes the kth derivative of a function $g:\mathbb{R} \to \mathbb{R}$. Define

$$\mathbf{w}_j(x_j, u_j) = \left(1, \left(\frac{u_j - x_j}{h_j}\right), \dots, \left(\frac{u_j - x_j}{h_j}\right)^\pi\right)^\top.$$

We consider the following kernel-weighted least-squares criterion to estimate m_j:

$$L(\mathbf{f}) = \int n^{-1} \sum_{i=1}^n \left[Y^i - \sum_{j=1}^d Z_j^i \mathbf{w}_j(x_j, X_j^i)^\top \mathbf{f}_j(x_j)\right]^2 K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i)\,\mathrm{d}\mathbf{x}, \tag{4.1}$$

where $\mathbf{f}^\top = (\mathbf{f}_1^\top, \dots, \mathbf{f}_d^\top)$ and $\mathbf{f}_j(x_j) = (f_{j,0}(x_j), \dots, f_{j,\pi}(x_j))^\top$ for functions $f_{j,k}:\mathbb{R} \to \mathbb{R}$. Let $\hat{\mathbf{m}}$ be the minimizer of $L(\mathbf{f})$. The proposed estimators of m_j are then $\hat{m}_{j,0}$ in $\hat{\mathbf{m}}$, and $\hat{m}_{j,k}$ in $\hat{\mathbf{m}}$ are estimators of $h^k m_j^{(k)}/k!$. We thus define the proposed estimators of $m_j^{(k)}$ by

$$\hat{m}_j^{(k)}(x_j) = k! h_j^{-k} \hat{m}_{j,k}(x_j), \qquad 0 \le k \le \pi,\ 1 \le j \le d.$$

The minimization of $L(\mathbf{f})$ at (4.1) is done over $\mathbf{f}$ with $L(\mathbf{f}) < \infty$. Define

$$\mathbf{v}(\mathbf{u}, \mathbf{z}; \mathbf{x})^\top = (\mathbf{w}_1(x_1, u_1)^\top z_1, \dots, \mathbf{w}_d(x_d, u_d)^\top z_d).$$

The expression in the bracket at (4.1) can then be written as $Y^i - \mathbf{v}(\mathbf{X}^i, \mathbf{Z}^i; \mathbf{x})^\top \mathbf{f}(\mathbf{x})$. We now redefine $\hat{\mathbf{M}}$ used in the previous two sections as

$$\hat{\mathbf{M}}(\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbf{v}(\mathbf{X}^i, \mathbf{Z}^i; \mathbf{x}) \mathbf{v}(\mathbf{X}^i, \mathbf{Z}^i; \mathbf{x})^\top K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i). \tag{4.2}$$

$L(\mathbf{f}) < \infty$ is then equivalent to $\int \mathbf{f}(\mathbf{x})^\top \hat{\mathbf{M}}(\mathbf{x}) \mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x} < \infty$ and minimizing $L(\mathbf{f})$ is equivalent to minimizing $\int [\tilde{\mathbf{m}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})]^\top \hat{\mathbf{M}}(\mathbf{x}) [\tilde{\mathbf{m}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})]\,\mathrm{d}\mathbf{x}$, where $\tilde{\mathbf{m}}(\mathbf{x})$ is redefined as

$$\tilde{\mathbf{m}}(\mathbf{x}) = \hat{\mathbf{M}}(\mathbf{x})^{-1} n^{-1} \sum_{i=1}^n \mathbf{v}(\mathbf{X}^i, \mathbf{Z}^i; \mathbf{x}) Y^i K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i). \tag{4.3}$$

The function space that arises in this general problem is the class of $(\pi + 1)d$-vectors of functions $\mathbf{f} = (f_{j,k})$ such that $\int \mathbf{f}(\mathbf{x})^\top \hat{\mathbf{M}}(\mathbf{x}) \mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x} < \infty$ and $f_{j,k}(\mathbf{x}) = g_{j,k}(x_j)$ for some functions $g_{j,k}:\mathbb{R} \to \mathbb{R}$, $1 \le j \le d$ and $0 \le k \le \pi$. We continue to denote the function space by $\mathcal{H}(\hat{\mathbf{M}})$, and its norm by $\|\cdot\|_{\hat{\mathbf{M}}}$. Thus,

$$\hat{\mathbf{m}} = \operatorname*{argmin}_{\mathbf{f} \in \mathcal{H}(\hat{\mathbf{M}})} \|\tilde{\mathbf{m}} - \mathbf{f}\|_{\hat{\mathbf{M}}}^2. \tag{4.4}$$

By considering the Gâteaux or Fréchet derivatives of the objective function $L(\mathbf{f})$ with respect to $\mathbf{f}$, the solution $\hat{\mathbf{m}}$ of the minimization problem (4.4) satisfies the following system of integral equations:

$$\mathbf{0} = \int \hat{\mathbf{M}}_j(\mathbf{x})^\top [\tilde{\mathbf{m}}(\mathbf{x}) - \hat{\mathbf{m}}(\mathbf{x})] \,\mathrm{d}\mathbf{x}_{-j}, \qquad 1 \le j \le d, \tag{4.5}$$

where $\mathbf{0}$ is the $(\pi+1)$-dimensional zero vector and $\hat{\mathbf{M}}_j$ are $(\pi+1)d \times (\pi+1)$ matrices defined by $\hat{\mathbf{M}} = \hat{\mathbf{M}}^\top = (\hat{\mathbf{M}}_1, \ldots, \hat{\mathbf{M}}_d)$. We write $\hat{\mathbf{m}}^\top = (\hat{\mathbf{m}}_1^\top, \ldots, \hat{\mathbf{m}}_d^\top)$. Define

$$\tilde{\mathbf{m}}_j(x_j) = \hat{\mathbf{\Psi}}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \mathbf{w}_j(x_j, X_j^i) K_{h_j}(x_j, X_j^i) Z_j^i Y^i,$$

$$\hat{\mathbf{\Psi}}_j(x_j) = n^{-1} \sum_{i=1}^n \mathbf{w}_j(x_j, X_j^i) \mathbf{w}_j(x_j, X_j^i)^\top K_{h_j}(x_j, X_j^i)(Z_j^i)^2,$$

$$\hat{\mathbf{\Psi}}_{jk}(x_j, x_k) = n^{-1} \sum_{i=1}^n \mathbf{w}_j(x_j, X_j^i) \mathbf{w}_k(x_k, X_k^i)^\top K_{h_j}(x_j, X_j^i) K_{h_k}(x_k, X_k^i) Z_j^i Z_k^i$$

for $k \ne j$. We then find that the system of $(\pi+1)$-dimensional equations (4.5) is equivalent to the following backfitting equations which update the estimators of m_j and their derivatives up to the πth order:

$$\hat{\mathbf{m}}_j(x_j) = \tilde{\mathbf{m}}_j(x_j) - \sum_{k=1,\ne j}^d \int \hat{\mathbf{\Psi}}_j(x_j)^{-1} \hat{\mathbf{\Psi}}_{jk}(x_j, x_k) \hat{\mathbf{m}}_k(x_k) \,\mathrm{d}x_k, \qquad 1 \le j \le d. \tag{4.6}$$

We want to emphasize again that the method with local polynomial fitting does not require computation of the full-dimensional estimator $\tilde{\mathbf{m}}(\mathbf{x})$ at (4.3). It only requires one- and two-dimensional smoothing to compute $\tilde{\mathbf{m}}_j$, $\hat{\mathbf{\Psi}}_j$ and $\hat{\mathbf{\Psi}}_{jk}$, and involves inversion of $\hat{\mathbf{\Psi}}_j$ only. Although $\hat{\mathbf{\Psi}}_j$ are $(\pi+1) \times (\pi+1)$ matrices, they are computed by means of one-dimensional local smoothing so that they do not suffer from sparsity of data in high dimensions. The marginal integration method, in contrast, requires the computation of $\tilde{\mathbf{m}}(\mathbf{x})$ and thus, in practice, the marginal integration method may break down in the case where d is large.

Backfitting algorithm. *With a set of initial estimates* $\hat{\mathbf{m}}_j^{[0]} = (\hat{m}_{j,0}, \ldots, \hat{m}_{j,\pi})^\top$, *we iterate for* $r = 1, 2, \ldots$ *the following process: for* $1 \le j \le d$,

$$\begin{aligned} \hat{\mathbf{m}}_j^{[r]}(x_j) = \tilde{\mathbf{m}}_j(x_j) &- \sum_{k=1}^{j-1} \int \hat{\mathbf{\Psi}}_j(x_j)^{-1} \hat{\mathbf{\Psi}}_{jk}(x_j, x_k) \hat{\mathbf{m}}_k^{[r]}(x_k) \,\mathrm{d}x_k \\ &- \sum_{k=j+1}^{d} \int \hat{\mathbf{\Psi}}_j(x_j)^{-1} \hat{\mathbf{\Psi}}_{jk}(x_j, x_k) \hat{\mathbf{m}}_k^{[r-1]}(x_k) \,\mathrm{d}x_k. \end{aligned} \tag{4.7}$$

In the following two theorems, we show that the backfitting algorithm (4.7) converges to $\hat{\mathbf{m}}_j, 1 \le j \le d$, at a geometric rate and that $\hat{\mathbf{m}}_j, 1 \le j \le d$, are jointly asymptotically normal. We give the results for the case where π, the order of local polynomial fitting, is odd. It is widely accepted that fitting odd orders of local polynomial is better than even orders. It also gives simpler formulas in the asymptotic expansion and requires a weaker smoothness condition on $E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})$. In fact, instead of (A6) in Section 3, we need the following assumption:

(A7$'$) The coefficient functions m_j are $(\pi+1)$-times continuously differentiable on $[0,1]$ and $E(Z_jZ_k|\mathbf{X}=\mathbf{x})$ is continuous on $[0,1]^d$ for all $1\le j,k\le d$.

To state the first theorem, we need to introduce the limit of the matrix $\hat{\mathbf{M}}(\mathbf{x})$. Note that $\hat{\mathbf{M}}(\mathbf{x})$ consists of $(\pi+1)\times(\pi+1)$ blocks

$$\hat{\mathbf{M}}_{j,k}(\mathbf{x}) \equiv n^{-1}\sum_{i=1}^n \mathbf{w}_j(x_j,X_j^i)\mathbf{w}_k(x_k,X_k^i)^\top Z_j^iZ_k^iK_{\mathbf{h}}(\mathbf{x},\mathbf{X}^i), \qquad 1\le j,k\le d.$$

For $j\ne k$, the matrices $\hat{\mathbf{M}}_{j,k}(\mathbf{x})$ are approximated by

$$E[\mathbf{w}_j(x_j,X_j)\mathbf{w}_k(x_k,X_k)^\top Z_j^iZ_k^iK_{\mathbf{h}}(\mathbf{x},\mathbf{X})] \simeq \boldsymbol{\mu}\boldsymbol{\mu}^\top E(Z_jZ_k|\mathbf{X}=\mathbf{x})p(\mathbf{x}),$$

where $\boldsymbol{\mu}=(\mu_\ell(K))^\top$ and $\mu_\ell(K)=\int u^\ell K(u)\,\mathrm{d}u$. On the other hand, for $j=k$,

$$\hat{\mathbf{M}}_{j,j}(\mathbf{x}) \simeq \mathbf{N}_1E(Z_j^2|\mathbf{X}=\mathbf{x})p(\mathbf{x}),$$

where $\mathbf{N}_1$ is a $(\pi+1)\times(\pi+1)$ matrix defined by $\mathbf{N}_1=(\mu_{\ell+\ell'}(K))$. Here, we adopt the convention that the indices of the matrix entries run from $(0,0)$ to (π,π). Thus, $\hat{\mathbf{M}}(\mathbf{x})$ is approximated by

$$\mathbf{M}(\mathbf{x}) \equiv p(\mathbf{x})[E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})\otimes(\boldsymbol{\mu}\boldsymbol{\mu}^\top)+\mathrm{diag}(E(Z_j^2|\mathbf{X}=\mathbf{x}))\otimes(\mathbf{N}_1-\boldsymbol{\mu}\boldsymbol{\mu}^\top)], \qquad (4.8)$$

where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{M}(\mathbf{x})$ is positive definite under the assumption (A1). To see this, note first that by (A1), the matrix $E(\mathbf{Z}\mathbf{Z}^\top|\mathbf{X}=\mathbf{x})\otimes(\boldsymbol{\mu}\boldsymbol{\mu}^\top)$ is nonnegative definite. Also, $E(Z_j^2|\mathbf{X}=\mathbf{x})$ are bounded away from zero on $[0,1]^d$ for all $1\le j\le d$. Furthermore, $\mathbf{N}_1-\boldsymbol{\mu}\boldsymbol{\mu}^\top$ is the variance–covariance matrix of $(1,U,\ldots,U^\pi)^\top$, where U is a random variable with density K. Since K is supported on a uncountable set, it follows that $\mathbf{N}_1-\boldsymbol{\mu}\boldsymbol{\mu}^\top$ is positive definite. The foregoing arguments show that the smallest eigenvalue of $\mathbf{M}(\mathbf{x})$ is bounded away from zero on $[0,1]^d$. Let $\mathcal{H}(\mathbf{M})$ be defined as $\mathcal{H}(\hat{\mathbf{M}})$ with $\hat{\mathbf{M}}$ being replaced by $\mathbf{M}$ and define its norm by $\|\mathbf{f}\|_{\mathbf{M}}^2=\int\mathbf{f}(\mathbf{x})^\top\mathbf{M}(\mathbf{x})\mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

Theorem 3. *Assume that* (A1)–(A5) *hold. Then, with probability tending to one, there exists a solution* $\{\hat{\mathbf{m}}_j\}_{j=1}^d$ *of the backfitting equation (4.6) that is unique. Furthermore, there exist constants* $0<\gamma<1$ *and* $0<C<\infty$ *such that, with probability tending to one,*

$$\begin{aligned}&\sum_{j=1}^d\int|\hat{\mathbf{m}}_j^{[r]}(x_j)-\hat{\mathbf{m}}_j(x_j)|^2p_j(x_j)\,\mathrm{d}x_j\\&\quad\le C\gamma^{2r}\sum_{j=1}^d\int[|\tilde{\mathbf{m}}_j(x_j)|^2+|\hat{\mathbf{m}}_j^{[0]}(x_j)|^2]p_j(x_j)\,\mathrm{d}x_j.\end{aligned}$$

In the next theorem, we give the asymptotic distribution of the proposed estimators. We define $\mathbf{m}(\mathbf{x})^\top = (\mathbf{m}_1(x_1)^\top, \ldots, \mathbf{m}_d(x_d)^\top)$, where

$$\mathbf{m}_j(x_j) = (m_j(x_j), h_j m_j^{(1)}(x_j)/1!, \ldots, h_j^\pi m_j^{(\pi)}(x_j)/\pi!)^\top. \tag{4.9}$$

For the bandwidths h_j, we assume that h_j is asymptotic to $c_j n^{-1/(2\pi+3)}$ for some constant $0 < c_j < \infty$. Define $\boldsymbol{\gamma} = (\mu_{\pi+1}(K), \ldots, \mu_{\pi+1+\pi}(K))^\top$ and a $(\pi+1) \times (\pi+1)$ matrix by $\mathbf{N}_2 = (\mu_{\ell+\ell'}(K^2))$. For $1 \le j \le d$, define $\boldsymbol{\beta}_j(x_j) = c_j^{\pi+1}\mathbf{N}_1^{-1}\boldsymbol{\gamma} m_j^{(\pi+1)}(x_j)/(\pi+1)!$ and

$$\mathbf{V}_j(x_j) = \frac{E[Z_j^2\sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j]}{c_j p_j(x_j)[E(Z_j^2|X_j = x_j)]^2}\mathbf{N}_1^{-1}\mathbf{N}_2\mathbf{N}_1^{-1}.$$

Theorem 4. *Assume that* (A1)–(A6) *and* (A7′) *hold, and that the bandwidths* h_j *are asymptotic to* $c_j n^{-1/(2\pi+3)}$ *for some constants* $0 < c_j < \infty$. *Then, for any* $\mathbf{x} \in (0,1)^d$, $n^{(\pi+1)/(2\pi+3)} \times [\hat{\mathbf{m}}_j(x_j) - \mathbf{m}_j(x_j)]$, $1 \le j \le d$, *are asymptotically independent and*

$$n^{(\pi+1)/(2\pi+3)}[\hat{\mathbf{m}}_j(x_j) - \mathbf{m}_j(x_j)] \xrightarrow{d} N(\boldsymbol{\beta}_j(x_j), \mathbf{V}_j(x_j)), \qquad 1 \le j \le d.$$

Theorem 4 not only gives the asymptotic distributions of the estimators of the coefficient functions m_j, but also those of their derivatives. Recall the definition of $\mathbf{m}_j$ at (4.9) and that $\hat{\mathbf{m}}_j(x_j) = (\hat{m}_j(x_j), h_j\hat{m}_j^{(1)}(x_j)/1!, \ldots, h_j^\pi \hat{m}_j^{(\pi)}(x_j)/\pi!)^\top$. Thus, the theorem implies that $n^{(\pi+1-k)/(2\pi+3)}[\hat{m}_j^{(k)}(x_j) - m_j^{(k)}(x_j)]$ is asymptotically normal with mean $k!c_j^{\pi+1-k}(\mathbf{N}_1^{-1}\boldsymbol{\gamma})_k \times m_j^{(\pi+1)}(x_j)/(\pi+1)!$ and variance

$$(k!)^2(\mathbf{N}_1^{-1}\mathbf{N}_2\mathbf{N}_1^{-1})_{kk}\frac{E[Z_j^2\sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j]}{c_j^{2k+1}p_j(x_j)[E(Z_j^2|X_j = x_j)]^2},$$

where, for a vector $\mathbf{a}$ and a matrix $\mathbf{B}$, $\mathbf{a}_k$ denotes the kth entry of $\mathbf{a}$ and $\mathbf{B}_{kk}$ denotes the kth diagonal entry of $\mathbf{B}$. In the case of local linear fitting $(\pi = 1)$, we have

$$(\mathbf{N}_1^{-1}\mathbf{N}_2\mathbf{N}_1^{-1})_{00} = \int K^2(u)\,\mathrm{d}u, \qquad (\mathbf{N}_1^{-1}\boldsymbol{\gamma})_0 = \int u^2 K(u)\,\mathrm{d}u.$$

Another implication of Theorem 4 is that the estimators $\hat{m}_j^{(k)}(x_j)$ for $0 \le k \le \pi$ have the oracle properties. Suppose that we know all other coefficient functions except m_j. In this case, we would estimate m_j and its derivatives up to order π by minimizing

$$n^{-1}\sum_{i=1}^n\left[Y^i - \sum_{k=1,\neq j}^d m_k(X_k^i)Z_k^i - Z_j^i\mathbf{w}_j(x_j,X_j^i)^\top\mathbf{f}_j(x_j)\right]^2 K_{h_j}(x_j,X_j^i)$$

over $\mathbf{f}_j$. It can be shown that the resulting estimators of $m_j^{(k)}(x_j)$ for $0 \le k \le \pi$ have the same asymptotic distributions as $\hat{m}_j^{(k)}(x_j)$ for $0 \le k \le \pi$.

The asymptotically optimal choices of the bandwidths h_j may be derived from Theorem 4. Let $c_j^{\pi+1} b_j(x_j)$ and $c_j^{-1}\tau_j(x_j)$ denote the asymptotic mean and the asymptotic variance of $n^{(\pi+1)/(2\pi+3)}[\hat{m}_j(x_j) - m_j(x_j)]$, respectively. That is,

$$\begin{aligned} b_j(x_j) &= (\mathbf{N}_1^{-1}\boldsymbol{\gamma})_0 m_j^{(\pi+1)}(x_j)/(\pi+1)!, \\ \tau_j(x_j) &= \frac{E[Z_j^2\sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j]}{p_j(x_j)[E(Z_j^2|X_j = x_j)]^2}(\mathbf{N}_1^{-1}\mathbf{N}_2\mathbf{N}_1^{-1})_{00}. \end{aligned}$$

The optimal choice of c_j which minimizes the asymptotic mean integrated squared error is then given by

$$c_j^{\mathrm{opt}} = \left[\frac{\int \tau_j(x_j)p_j(x_j)\,\mathrm{d}x_j}{2(\pi+1)\int b_j(x_j)^2 p_j(x_j)\,\mathrm{d}x_j}\right]^{1/(2\pi+3)}. \tag{4.10}$$

This formula for the optimal bandwidth involves unknown quantities. We may get a rule-of-thumb bandwidth selector by fitting polynomial regression models, as in Yang *et al.* [24], to estimate the unknown quantities in the formula for c_j^{opt}; see Section 6, where we employ this approach to analyze Boston Housing Data. Alternatively, we may adopt the approach of Mammen and Park [18] to obtain more sophisticated bandwidth selectors.

5. Numerical properties

We investigated the finite-sample properties of the proposed estimators in comparison with the marginal integration method studied in Yang *et al.* [24]. We considered the case where local linear smoothing ($\pi = 1$) is employed. The simulation study was done in two settings, one in a low-dimensional case ($d = 3$) and the other in a high-dimensional case ($d = 10$).

In the first case, we generated the data $(\mathbf{X}^i, \mathbf{Z}^i, Y^i)$ from the model: $Y = m_1(X_1)Z_1 + m_2(X_2)Z_2 + m_3(X_3)Z_3 + \sigma(\mathbf{X},\mathbf{Z})\varepsilon$, where $Z_1 \equiv 1$ and

$$\sigma(\mathbf{x},\mathbf{z}) = \frac{1}{2} + \frac{z_2^2 + z_3^2}{1 + z_2^2 + z_3^2}\exp\left(-2 + \frac{x_1 + x_2}{2}\right). \tag{5.1}$$

The vector $\mathbf{X} = (X_1, X_2, X_3)$ was generated from the uniform distribution over the unit cube $(0,1)^3$, and the covariate vector (Z_2, Z_3) was generated from the bivariate normal with mean $(0,0)$ and covariance matrix $\left(\begin{smallmatrix} 1 & 0.5 \\ 0.5 & 1 \end{smallmatrix}\right)$. The vectors $\mathbf{X}$ and $\mathbf{Z}$ were independent, and the error term ε was generated from the standard normal distribution, independently of $(\mathbf{X},\mathbf{Z})$. This model was also considered in Yang *et al.* [24]. We took $m_1(x) = 1 + \mathrm{e}^{2x-1}$, $m_2(x) = \cos(2\pi x)$ and $m_3(x) = x^2$.

In the second case, where $d = 10$, we took the same variance function $\sigma^2(\mathbf{x},\mathbf{z})$ as in (5.1), for the sake of simplicity. Thus, $\sigma^2(\mathbf{x},\mathbf{z})$ did not depend on (x_j, z_j) for $4 \le j \le 10$. The extra covariates X_j for $4 \le j \le 10$ were generated from the uniform distribution over $(0,1)^7$ independently of (X_1, X_2, X_3), and Z_j for $4 \le j \le 10$ were generated from the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$, the identity matrix, independently of (Z_2, Z_3) and of $\mathbf{X}$. We chose $m_j(x) = x^2$ for $4 \le j \le 10$.

We used the Epanechnikov kernel $K(u) = (3/4)(1 - u^2)I[-1,1](u)$ and the optimal bandwidths $h_j^{\mathrm{sbf}} = c_j^{\mathrm{opt}} n^{-1/5}$, where c_j^{opt} are given at (4.10). This was for the proposed estimator. For the marginal integration method, the estimator $\hat{m}_j^{\mathrm{mi}}$ of the jth coefficient function m_j that we investigated was

$$\hat{m}_j^{\mathrm{mi}}(x_j) = n^{-1}\sum_{i=1}^{n}\hat{\theta}_j(X_1^i,\ldots,X_{j-1}^i,x_j,X_{j+1}^i,\ldots,X_j^n),$$

where $\hat{\theta}_j(\mathbf{x})$ was the $[(j-1)(\pi+1)+1]$st entry of $\tilde{\mathbf{m}}(\mathbf{x})$ defined at (4.3), but K_{h_k} (for $k \neq j$) in the definition of $\tilde{\mathbf{m}}(\mathbf{x})$ was replaced by L_{b_k}. Note that, for the marginal integration method, in the estimation of the jth coefficient function, we may choose another kernel L and need to use other bandwidths b_k, different from h_j, for the directions of $x_k (k \neq j)$ not of interest. We took $L = K$ and $b_k = c(\log n)^{-1}h_j^{\mathrm{mi}}$ for all directions $k \neq j$, where h_j^{mi} is the optimal bandwidth for the marginal integration method, obtained similarly as the one for the proposed method at (4.10), and c was a constant multiplier for which we tried four values, $c = 1, 3, 5, 10$.

We used $\tilde{\mathbf{m}}_j$ defined in Section 4 as the initial estimates $\hat{\mathbf{m}}_j^{[0]}$ for the proposed method. The backfitting algorithm converged very fast. We took

$$\sqrt{\sum_{j=1}^{d}\int[\hat{m}_j^{[r-1]}(x_j) - \hat{m}_j^{[r]}(x_j)]^2\,\mathrm{d}x_j} \leq 10^{-11}$$

as a criterion for the convergence. With this criterion, the backfitting algorithm converged within 11 iterations in all cases. The average number of iterations was 6.5 from the 500 replications. In a preliminary numerical study with the marginal integration method, we found that inverting the matrix $\hat{\mathbf{M}}(\mathbf{x})$ often caused numerical instability of the estimates, even for the low-dimensional case where $d = 3$. This reflects the curse of dimensionality that the marginal integration suffers from. Thus, we actually computed a 'ridged' version of $\hat{m}_j^{\mathrm{mi}}$ by adding n^{-2} to the diagonal entries of the matrix $\hat{\mathbf{M}}(\mathbf{x})$. The same modification was also made in the numerical study of Yang *et al.* [24].

Table 1 shows the results for the case $d = 3$, based on 500 data sets with sizes $n = 100$ and 400. The table provides the mean integrated squared errors (MISE) of the estimators of each coefficient function m_j, defined by

$$\begin{aligned}\mathrm{MISE}_j(\bar{m}_j) &= \int E[\bar{m}_j(x_j) - m_j(x_j)]^2\,\mathrm{d}x_j\\ &= \int[E\bar{m}_j(x_j) - m_j(x_j)]^2\,\mathrm{d}x_j + \int[\bar{m}_j(x_j) - E\bar{m}_j(x_j)]^2\,\mathrm{d}x_j\\ &\overset{\mathrm{let}}{=} \mathrm{ISB}_j(\bar{m}_j) + \mathrm{IV}_j(\bar{m}_j)\end{aligned}$$

for an estimator $\bar{m}_j$. It also gives the integrated squared bias (ISB) and the integrated variance (IV). The results suggest that the proposed method gives better performance in terms of $\mathrm{MISE}_{\mathrm{tot}} = \sum_{j=1}^{3}\mathrm{MISE}_j$. When $n = 100$, the sum of MISE_j of $\hat{m}_j$ equals 0.7621,

Table 1. The mean integrated squared errors (MISE), the integrated squared biases (ISB) and the integrated variances (IV) of the marginal integration estimators (MI) and the proposed estimators (SBF) when $d = 3$ (the constant c for MI is the multiplier c in the formula $b_k = c(\log n)^{-1}h_j^{\mathrm{mi}}$, where b_k is the secondary bandwidth applied to the direction of x_k, $k \neq j$, in the estimation of m_j)

Sample size	Coefficient function		MI				SBF
			$c = 1$	$c = 3$	$c = 5$	$c = 10$	
$n = 100$	m_1	MISE	0.1190	0.1140	0.1158	0.1151	0.1496
		ISB	0.0174	0.0150	0.0147	0.0145	0.0019
		IV	0.1016	0.0990	0.1011	0.1006	0.1476
	m_2	MISE	0.6354	0.5738	0.5795	0.5826	0.3613
		ISB	0.4089	0.3502	0.3465	0.3461	0.0484
		IV	0.2265	0.2236	0.2330	0.2364	0.3129
	m_3	MISE	0.1873	0.2218	0.2255	0.2259	0.2512
		ISB	0.0057	0.0056	0.0056	0.0056	0.0017
		IV	0.1816	0.2163	0.2200	0.2203	0.2495
$n = 400$	m_1	MISE	0.0347	0.0332	0.0365	0.0363	0.0415
		ISB	0.0092	0.0087	0.0087	0.0086	0.0005
		IV	0.0255	0.0245	0.0279	0.0277	0.0410
	m_2	MISE	0.2648	0.2815	0.2872	0.2894	0.1244
		ISB	0.2126	0.2227	0.2248	0.2257	0.0199
		IV	0.0521	0.0588	0.0624	0.0637	0.1045
	m_3	MISE	0.0478	0.0576	0.0610	0.0620	0.0810
		ISB	0.0050	0.0046	0.0046	0.0047	0.0008
		IV	0.0428	0.0529	0.0564	0.0573	0.0802

while those of the marginal integration method are $0.9417, 0.9096, 0.9208, 0.9236$ for $c = 1, 3, 5, 10$, respectively. In the case where $n = 400$, $\mathrm{MISE}_{\mathrm{tot}} = 0.2469$ for the proposed method, while it equals $0.3473, 0.3723, 0.3847, 0.3877$ for the marginal integration method.

According to Table 1, the performance of the marginal integration method appears not to be sensitive to the choice of the secondary bandwidth b_k. However, this is true only when we use the optimal bandwidth h_j^{mi}. In fact, we found that the performance depended crucially on the choice b_k when other choices of h_j were used. As an example, we report in Table 2 the results when one uses $h_j = h_j^{\mathrm{mi}}/3$ instead of $h_j = h_j^{\mathrm{mi}}$. In the latter case, the sum of MISE_j ranges from 0.8001 to 2.7453 when $n = 100$, and from 0.2291 to 2.1080 when $n = 400$, for those four values of c. One interesting thing to note is that the ISB of the marginal integration increases drastically as c decreases. The main lesson here is that the choice of the secondary bandwidths b_k for the marginal integration method is as important as the choice of h_j.

The finite-sample results in Table 1 show some discrepancy with the asymptotics for the functions m_1 and m_3. Asymptotically, if the optimal bandwidth is used, then the IV is four times as large as the ISB. In general, finite-sample properties do not always match

Table 2. The mean integrated squared errors (MISE), the integrated squared biases (ISB) and the integrated variances (IV) of MI when $h_j = h_j^{\mathrm{mi}}/3$ was used (the constant c is the multiplier c in the formula $b_k = c(\log n)^{-1} h_j^{\mathrm{mi}}$, where b_k is the secondary bandwidth applied to the direction of x_k, $k \neq j$, in the estimation of m_j)

Sample size	Coefficient function		MI			
			$c=1$	$c=3$	$c=5$	$c=10$
$n=100$	m_1	MISE	1.5109	0.2664	0.1822	0.1737
		ISB	1.4327	0.0096	0.0011	0.0012
		IV	0.0782	0.2568	0.1812	0.1725
	m_2	MISE	0.6611	0.4578	0.3459	0.3576
		ISB	0.3095	0.0340	0.0338	0.0313
		IV	0.3516	0.4238	0.3121	0.3263
	m_3	MISE	0.5733	0.2743	0.2720	0.3012
		ISB	0.0217	0.0013	0.0014	0.0015
		IV	0.5516	0.2730	0.2706	0.2997
$n=400$	m_1	MISE	1.4539	0.0891	0.0465	0.0465
		ISB	1.4177	0.0032	0.0004	0.0003
		IV	0.0362	0.0859	0.0461	0.0462
	m_2	MISE	0.3554	0.1596	0.1109	0.1129
		ISB	0.2359	0.0139	0.0154	0.0160
		IV	0.1195	0.1457	0.0955	0.0969
	m_3	MISE	0.2987	0.0702	0.0717	0.0856
		ISB	0.0188	0.0005	0.0006	0.0007
		IV	0.2799	0.0697	0.0711	0.0849

with asymptotics. One possible reason for the discrepancy in this particular setting is that the coefficient functions m_1 and m_3 are far simpler than the complexity brought by the noise level, so the proposed method easily catches the structure with less bias. This seems not to be the case with the marginal integration, however. For the marginal integration, the secondary bandwidths b_k interact with the primary bandwidth h_j for the bias and variance performance, as discussed in the previous paragraph.

Table 3 shows the results for the case $d = 10$. Here, for the marginal integration, we report only the results when $c = 5$ which gave the best performance. In fact, the marginal integration got worse very quickly as c decreased from $c = 5$. For example, we found the total MISE, $\sum_{j=1}^{10} \mathrm{MISE}_j$, was 3.4996 when $c = 3$ and was 6.1834 when $c = 1$, in the case where $n = 400$. Note that the value equals 0.6440 when $c = 5$ and $n = 400$, as reported in Table 3. For the proposed method, it equals 0.5136.

6. Analysis of Boston Housing Data

The data consist of fourteen variables, among which one is response and the other thirteen are predictors. There are 506 observations from 506 tracts in the Boston area; see Harrison

Table 3. The mean integrated squared errors (MISE), the integrated squared biases (ISB) and the integrated variances (IV) of the marginal integration estimators (MI) and the proposed estimators (SBF) when $d = 10$

Sample size	Coefficient function	MI			SBF		
		MISE	ISB	IV	MISE	ISB	IV
$n = 100$	m_1	0.2533	0.1242	0.1291	0.1904	0.0046	0.1858
	m_2	0.7284	0.4353	0.2931	0.4357	0.0605	0.3752
	m_3	0.2622	0.0059	0.2563	0.3042	0.0024	0.3018
	m_4	0.1303	0.0054	0.1249	0.1404	0.0022	0.1382
	m_5	0.1351	0.0060	0.1291	0.1489	0.0011	0.1478
	m_6	0.1336	0.0055	0.1281	0.1509	0.0019	0.1490
	m_7	0.1345	0.0054	0.1291	0.1677	0.0019	0.1658
	m_8	0.1228	0.0053	0.1175	0.1482	0.0019	0.1463
	m_9	0.1428	0.0071	0.1357	0.1707	0.0009	0.1698
	m_{10}	0.1270	0.0059	0.1211	0.1528	0.0014	0.1514
$n = 400$	m_1	0.0505	0.0115	0.0390	0.0457	0.0008	0.0449
	m_2	0.2999	0.2223	0.0776	0.1264	0.0196	0.1068
	m_3	0.0642	0.0054	0.0588	0.0893	0.0004	0.0889
	m_4	0.0324	0.0048	0.0276	0.0379	0.0005	0.0374
	m_5	0.0358	0.0054	0.0304	0.0355	0.0010	0.0345
	m_6	0.0369	0.0040	0.0329	0.0331	0.0004	0.0327
	m_7	0.0300	0.0044	0.0256	0.0370	0.0009	0.0361
	m_8	0.0319	0.0043	0.0276	0.0368	0.0006	0.0362
	m_9	0.0321	0.0052	0.0269	0.0364	0.0009	0.0355
	m_{10}	0.0303	0.0046	0.0257	0.0355	0.0006	0.0349

and Rubinfeld [10] for details about the data set. The data set has been analyzed by Fan and Huang [7] and Wang and Yang [22], among others. The former fitted the data using a partially linear functional coefficient model where all coefficient functions in the nonparametric part are functions of a single variable. The latter considered an additive regression model. Here, we apply the varying coefficient model (1.1) to fit the data using the proposed method. We take the variable MEDV (median value of owner-occupied homes in \$1000's) as the response variable Y. We consider five variables as covariates X_j or Z_j. They are CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property tax rate per \$10 000), PTRATIO (pupil–teacher ratio by town) and LSTAT (percentage of lower income status of the population). As in Wang and Yang [22], we take logarithmic transformation for TAX and LSTAT to remove sparse areas in the domains of these variables.

We want to find a varying coefficient model that fits the data set well. Since LSTAT can be a good explanatory variable that determines the overall level of the housing price, we consider models of the form

$$\mathrm{MEDV} = m_1(\log(\mathrm{LSTAT})) + m_2(X_2)Z_2 + m_3(X_3)Z_3 + (\mathrm{noise}). \tag{6.1}$$

Table 4. Relative squared prediction errors obtained from fitting 12 varying coefficient models with the Boston Housing Data

Model no.	Covariates				Relative squared prediction error
	X_2	Z_2	X_3	Z_3	
1	CRIM	RM	TAX	PTRATIO	0.3514
2	CRIM	RM	PTRATIO	TAX	N/A
3	CRIM	TAX	RM	PTRATIO	0.2700
4	CRIM	TAX	PTRATIO	RM	0.2688
5	CRIM	PTRATIO	RM	TAX	0.4390
6	CRIM	PTRATIO	TAX	RM	0.4757
7	RM	CRIM	TAX	PTRATIO	0.3010
8	RM	CRIM	PTRATIO	TAX	*0.2412*
9	RM	TAX	PTRATIO	CRIM	N/A
10	RM	PTRATIO	TAX	CRIM	N/A
11	TAX	CRIM	PTRATIO	RM	N/A
12	TAX	RM	PTRATIO	CRIM	N/A

A general question is which variables should be the model covariates Z_j and which should take the role of X_j. This may be obvious for some data sets, but it is not so clear for the Boston Housing Data. Thus, we fitted all possible models and chose the one that best fitted the data. In general, we do not suggest employing the all-possible-models approach since it can get out of control quickly as the number of variables increases, and it induces a certain arbitrariness in the choice. For the Boston Housing Data, there are only twelve varying coefficient models of the form (6.1), listed in Table 4, and all models are interpretable. If the number of variables is large, then we suggest first choosing a set of model covariates Z_j among all covariates by fitting parametric linear models and using a variable selection technique, and then picking one as X_j for each Z_j from the remaining variables based on a criterion such as RSPE (which is defined later).

We employed local linear smoothing in implementing the proposed method and used the Epanechnikov kernel. For the bandwidths h_j, we chose to use a rule-of-thumb method that we describe below. Note that the unknowns in the expression of the optimal bandwidth at (4.10) are $A_j = \int m_j''(x_j)^2 p_j(x_j)\,\mathrm{d}x_j$, $B_j(x_j) = E[Z_j^2 \sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j]$ and $C_j(x_j) = E(Z_j^2|X_j = x_j)$. The second derivative of m_j in A_j can be estimated by fitting a cubic polynomial regression model. This gives $\hat{A}_j = n^{-1}\sum_{i=1}^n (2\hat{\alpha}_{j,2} + 6\hat{\alpha}_{j,3}X_j^i)^2$, where $\hat{\alpha}_{j,k}$ are the least-squares estimators that minimize

$$\sum_{i=1}^n \left[Y^i - \sum_{j=1}^d (\alpha_{j,0} + \alpha_{j,1}X_j^i + \alpha_{j,2}X_j^{i2} + \alpha_{j,3}X_j^{i3})Z_j^i \right]^2 .$$

Here, we take $Z_1^i \equiv 1$. The conditional means, B_j and C_j, can be estimated by fitting linear regression models. Since $E[Z_j^2\sigma^2(\mathbf{X},\mathbf{Z})|X_j = x_j] = E[Z_j^2(Y - m(\mathbf{X},\mathbf{Z}))^2|X_j = x_j]$, the conditional mean B_j is estimated by $\hat{B}_j(x_j) = \hat{\beta}_{j,0} + \hat{\beta}_{j,1}x_j$, where $\hat{\beta}_{j,0}$ and $\hat{\beta}_{j,1}$ minimize

$$\sum_{i=1}^{n} \left[Z_j^{i2} \left(Y^i - \sum_{k=1}^{d} (\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1}X_k^i + \hat{\alpha}_{k,2}X_k^{i2} + \hat{\alpha}_{k,3}X_k^{i3}) Z_k^i \right)^2 - \beta_{j,0} - \beta_{j,1}X_j^i \right]^2 .$$

Similarly, C_j for $j = 2, 3$ are estimated by $\hat{C}_j(x_j) = \hat{\gamma}_{j,0} + \hat{\gamma}_{j,1}x_j$, where $\hat{\gamma}_{j,0}$ and $\hat{\gamma}_{j,1}$ minimize $\sum_{i=1}^{n}(Z_j^{i2} - \gamma_{j,0} - \gamma_{j,1}X_j^i)^2$. Note that $C_1 \equiv 1$.

We split the data set into two parts, one for estimation of the models and the other for assessment of the estimated models. We selected 100 tracts for the model assessment out of 506 distributed in 92 towns. This was done in a manner that would lead to more selections in a town with a larger number of tracts. We fitted the twelve varying coefficient models using the data for the remaining 406 tracts and made out-of-sample predictions with the data for the selected 100 tracts. We calculated their relative squared prediction errors,

$$\mathrm{RSPE} = \frac{\sum_{i=1}^{100}[\mathrm{MEDV}^i - \hat{m}_1(\log(\mathrm{LSTAT}^i)) - \hat{m}_2(X_2^i)Z_2^i - \hat{m}_3(X_3^i)Z_3^i]^2}{\sum_{i=1}^{100}[\mathrm{MEDV}^i - \overline{\mathrm{MEDV}}]^2},$$

where $\hat{m}_j$ for $j = 1, 2, 3$ were constructed by using the data for the 406 remaining tracts.

Table 4 reports the results. In the table, we do not provide the values of RSPE for the models numbered 2, 9, 10, 11 and 12. In the preliminary fitting of these models taking X_j and Z_j as specified, we found that they produced extremely large residuals for some of the observations that corresponded to PTRATIO $= 20.2$ or TAX $= 666$. This resulted in a negative value of $\hat{B}_j(x_j)$ for a certain range of x_j and, as a consequence, produced a negative estimate of $\int \tau_j(x_j)p_j(x_j)\,\mathrm{d}x_j$ in the bandwidth formula (4.10). Since these five models do not explain MEDV well as a function of the covariates and would give a large value of RSPE when fitted, we excluded them from further analysis.

According to the table, the model with the smallest RSPE is

$$\mathrm{MEDV} = m_1(\log(\mathrm{LSTAT})) + m_2(\mathrm{RM})\,\mathrm{CRIM} + m_3(\mathrm{PTRATIO})\log(\mathrm{TAX}) + (\text{noise}). \tag{6.2}$$

Figure 1 depicts the estimated coefficient functions $\hat{m}_1, \hat{m}_2$ and $\hat{m}_3$. It also plots the actual values of MEDV and their predicted values according to the estimated model from (6.2). The prediction was made for those 100 tracts that were not used in estimating the model. The estimated curve $\hat{m}_1$ indicates that a high percentage of lower income status decreases the prices of homes. The estimated curve $\hat{m}_2$ suggests that for towns with higher or lower average numbers of rooms per dwelling, the crime rate is less influential on the prices of homes. Finally, from the estimated curve $\hat{m}_3$, we see that if the pupil–teacher ratio gets higher, then the prices of homes increase less rapidly as the property tax rate increases. The curve $\hat{m}_3$ looks somewhat rigid. The reason for this is that the variable PTRATIO does not really take values on a continuous scale since it is the pupil–teacher ratio by town, so that all tracts in a town have the same value of PTRATIO. Furthermore, some towns share the same value with others. For example, the 132 tracts (out of 506) associated with the 15 towns in the city of Boston have the same value, 20.2.

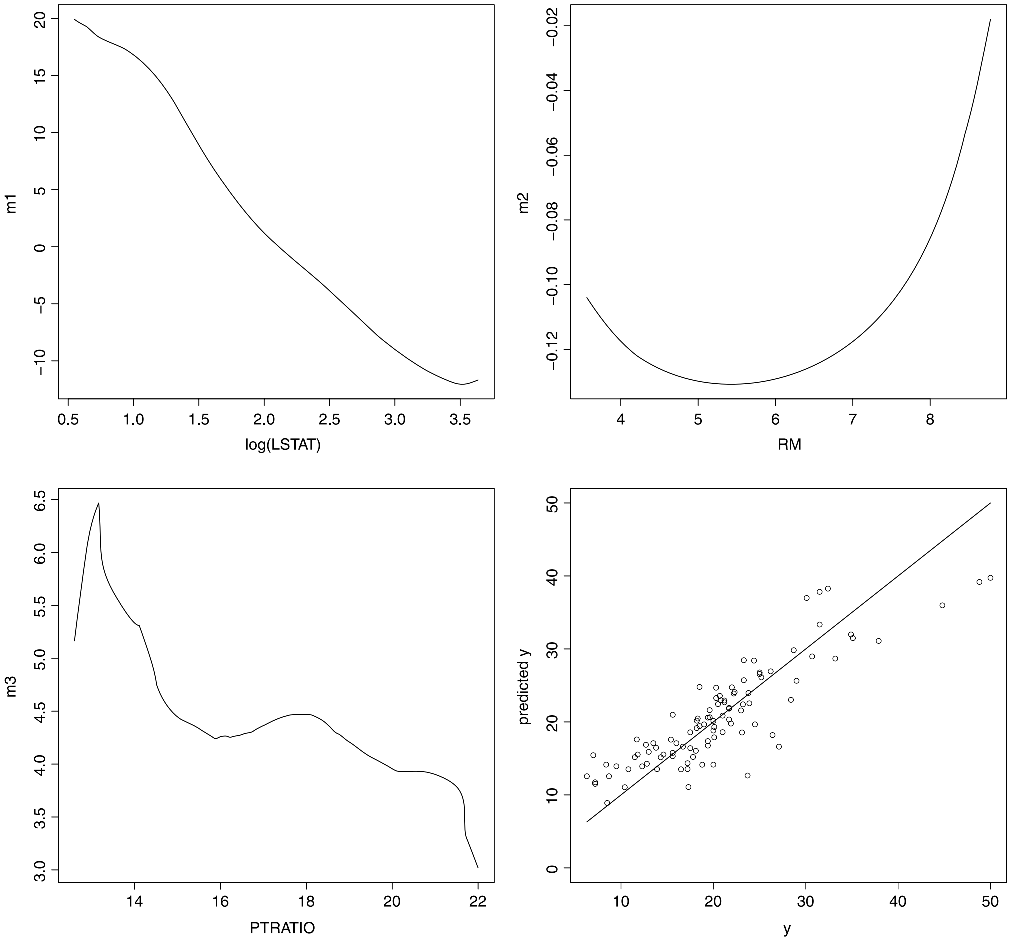

Figure 1. For the final model (6.2), the upper-left, upper-right and lower-left panels depict the estimated coefficient functions $\hat{m}_1$, $\hat{m}_2$ and $\hat{m}_3$, respectively, and the lower-right panel exhibits plots of the observed values Y^i versus their predicted values $\hat{Y}^i$.

Appendix: Technical details

A.1. Proof of Theorem 1

We prove that there exists a constant $0 < \gamma < 1$ such that $\|\hat{Q}\| < \gamma$ with probability tending to one. Let $\mathcal{H}_j(\mathbf{M})$ be defined as $\mathcal{H}_j(\hat{\mathbf{M}})$ with $\hat{\mathbf{M}}$ being replaced by $\mathbf{M}$. Let p_j and p_{jk} denote the marginal densities of X_j and (X_j, X_k), respectively. Define

$$q_j(x_j) = E(Z_j^2 | X_j = x_j) p_j(x_j), \tag{A.1}$$

$$q_{jk}(x_j,x_k)=E(Z_jZ_k|X_j=x_j,X_k=x_k)p_{jk}(x_j,x_k),\qquad k\neq j. \tag{A.2}$$

For $\mathbf{f}_j\in\mathcal{H}_j(\mathbf{M})$,

$$\|\mathbf{f}_j\|_{\mathbf{M}}^2=\int\mathbf{f}_j(\mathbf{x})^\top\mathbf{M}(\mathbf{x})\mathbf{f}_j(\mathbf{x})\,\mathrm{d}\mathbf{x}=\int f_j(x_j)^2q_j(x_j)\,\mathrm{d}x_j. \tag{A.3}$$

The equality (A.3) follows from the identity

$$\int E(Z_j^2|\mathbf{X}=\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j}=E(Z_j^2|X_j=x_j)p_j(x_j).$$

From (A.3) and Hölder's inequality, it follows that, for $\mathbf{f}\in\mathcal{H}(\mathbf{M})$,

$$\begin{aligned}
&\|(\hat{Q}_j-Q_j)\mathbf{f}\|_{\mathbf{M}}\\
&\quad=\left[\int\left(\sum_{k=1,\neq j}\int\left[\frac{\hat{q}_{jk}(x_j,x_k)}{\hat{q}_j(x_j)}-\frac{q_{jk}(x_j,x_k)}{q_j(x_j)}\right]f_k(x_k)\,\mathrm{d}x_k\right)^2q_j(x_j)\,\mathrm{d}x_j\right]^{1/2}\\
&\quad\le\sum_{k=1,\neq j}\left[\int\left(\frac{\hat{q}_{jk}(x_j,x_k)}{\hat{q}_j(x_j)q_k(x_k)}-\frac{q_{jk}(x_j,x_k)}{q_j(x_j)q_k(x_k)}\right)^2q_j(x_j)q_k(x_k)\,\mathrm{d}x_j\,\mathrm{d}x_k\right]^{1/2}\\
&\qquad\qquad\times\left[\int f_k(x_k)^2q_k(x_k)\,\mathrm{d}x_k\right]^{1/2}\\
&\quad\le\mathrm{o}_p(1)\sum_{k=1,\neq j}\|\mathbf{f}_k\|_{\mathbf{M}}.
\end{aligned}$$

Since $\|Q_j\|=1$, this proves that $\|\hat{Q}_j\|\le C_1$ with probability tending to one for some constant $0<C_1<\infty$. Define $Q=Q_d\cdots Q_1$. Then,

$$\|\hat{Q}-Q\|=\left\|\sum_{k=0}^{d-1}Q_d\cdots Q_{d-k+1}(\hat{Q}_{d-k}-Q_{d-k})\hat{Q}_{d-k-1}\cdots\hat{Q}_1\right\|=\mathrm{o}_p(1),$$

where we interpret both Q_{d+1} and $\hat{Q}_0$ as the zero operator. From (A1), (A3) and (A.3), the projection operators $\Pi_j:\mathcal{H}_k(\mathbf{M})\to\mathcal{H}_j(\mathbf{M})$ for all $1\le j\neq k\le d$ are Hilbert–Schmidt. By applying parts B, C and D of Proposition A.4.2 of Bickel, Klaassen, Ritov and Wellner [1], we find that $\|Q\|<1$. This shows that there exists a constant $0<\gamma<1$ such that $\|\hat{Q}\|<\gamma$ with probability tending to one.

To complete the proof of Theorem 1, it follows from (3.2) that with probability tending to one,

$$\|\hat{\mathbf{m}}^{[r]}-\hat{\mathbf{m}}\|_{\mathbf{M}}=\left\|\sum_{s=r}^{\infty}\hat{Q}^s\hat{\mathbf{r}}+\hat{Q}^r\hat{\mathbf{m}}^{[0]}\right\|_{\mathbf{M}}\le\gamma^r\left(\|\hat{\mathbf{r}}\|_{\mathbf{M}}\frac{1}{1-\gamma}+\|\hat{\mathbf{m}}^{[0]}\|_{\mathbf{M}}\right).$$

By (3.3) and the fact that $\|\hat{Q}_j\| \le C_1$ with probability tending to one, there exists a constant $0 < C_2 < \infty$ such that with probability tending to one,

$$\|\hat{\mathbf{r}}\|_{\mathbf{M}} \le C_2 \sum_{j=1}^{d} \biggl[\int \tilde{m}_j(x_j)^2 q_j(x_j)\,\mathrm{d}x_j\biggr]^{1/2}.$$

This completes the proof of Theorem 1.

A.2. Proof of Theorem 2

We will prove that for each $\mathbf{x} \in (0,1)^d$,

$$\hat{m}_j^A(x_j) = \tilde{m}_j^A(x_j) + \mathrm{o}_p(n^{-2/5}) \qquad \text{for } 1 \le j \le d, \tag{A.4}$$

$$\hat{\mathbf{m}}^B(\mathbf{x}) = \mathbf{m}(\mathbf{x}) + \boldsymbol{\beta}(\mathbf{x})n^{-2/5} + \mathrm{o}_p(n^{-2/5}). \tag{A.5}$$

Proof of (A.4). Note that $\hat{\mathbf{m}}^A = \sum_{s=0}^{\infty} \hat{Q}^s \hat{\mathbf{r}}^A$, where

$$\hat{\mathbf{r}}^A = (I - \hat{Q})\tilde{\mathbf{m}}^A = \tilde{\mathbf{m}}_d^A + \hat{Q}_d \tilde{\mathbf{m}}_{d-1}^A + \cdots + \hat{Q}_d \cdots \hat{Q}_2 \tilde{\mathbf{m}}_1^A \tag{A.6}$$

and $\tilde{\mathbf{m}}_j^A(\mathbf{x}) = (0,\ldots,0,\tilde{m}_j^A(x_j),0,\ldots,0)^\top$. From formulas (2.6)–(2.8), it follows that

$$\hat{Q}_d \cdots \hat{Q}_{j+1} \tilde{\mathbf{m}}_j^A(\mathbf{x}) = (0,\ldots,0,\tilde{m}_j^A(x_j),\tilde{g}_{j+1}(x_{j+1}),\ldots,\tilde{g}_d(x_d))^\top, \qquad 2 \le j \le d,$$

for some random functions $\tilde{g}_k : \mathbb{R} \to \mathbb{R}$, $j+1 \le k \le d$, where the first $j-1$ entries of the vector on the right-hand side of the equation are zero. This implies that

$$\hat{\mathbf{r}}^A(\mathbf{x}) = (\tilde{m}_1^A(x_1), \hat{g}_2(x_2), \ldots, \hat{g}_d(x_d))^\top, \tag{A.7}$$

where $\hat{g}_k$ for $2 \le k \le d$ are random functions from $\mathbb{R}$ to $\mathbb{R}$. If we prove that

$$\sup_{\mathbf{x}\in[0,1]^d} \Biggl|\sum_{s=1}^{\infty} \hat{Q}^s \hat{\mathbf{r}}^A(\mathbf{x})\Biggr| = \mathrm{o}_p(n^{-2/5}), \tag{A.8}$$

then (A.7) implies (A.4) for the case $j = 1$. By exchanging the entries of $\tilde{\mathbf{m}}^A$, we can see that (A.4) also holds for $j \ge 2$.

To prove (A.8), it suffices to show that

$$\sup_{\mathbf{x}\in[0,1]^d} |\hat{Q}\hat{\mathbf{r}}^A(\mathbf{x})| = \mathrm{o}_p(n^{-2/5}), \tag{A.9}$$

$$\Biggl\|\sum_{s=1}^{\infty} \hat{Q}^s \hat{\mathbf{r}}^A\Biggr\|_{\mathbf{M}} = \mathrm{o}_p(n^{-2/5}). \tag{A.10}$$

To see this, note that from (2.6) and (2.7), we have, for $\mathbf{f} = (f_1,\ldots,f_d)^\top \in \mathcal{H}(\hat{\mathbf{M}})$,

$$\hat{Q}_j \mathbf{f}(\mathbf{x}) = (f_1(x_1),\ldots,f_{j-1}(x_{j-1}), f_j^*(x_j), f_{j+1}(x_{j+1}),\ldots,f_d(x_d))^\top, \tag{A.11}$$

where $f_j^*(x_j) = -\sum_{k=1,\neq j}^d \int f_k(x_k)\frac{\hat{q}_{jk}(x_j,x_k)}{\hat{q}_j(x_j)}\,\mathrm{d}x_k$. Thus, there exists a constant $0 < C < \infty$ such that with probability tending to one,

$$\sup_{\mathbf{x}\in[0,1]^d}\left|\sum_{s=2}^{\infty}\hat{Q}^s\hat{\mathbf{r}}^A(\mathbf{x})\right| = \sup_{\mathbf{x}\in[0,1]^d}\left|\hat{Q}\sum_{s=1}^{\infty}\hat{Q}^s\hat{\mathbf{r}}^A(\mathbf{x})\right| \le C\left\|\sum_{s=1}^{\infty}\hat{Q}^s\hat{\mathbf{r}}^A\right\|_{\mathbf{M}}.$$

We prove (A.9) and (A.10). From standard kernel theory, we can prove that for all $k \neq j$,

$$\sup_{x_k\in[0,1]}\left|\int \tilde{m}_j^A(x_j)\frac{\hat{q}_{jk}(x_j,x_k)}{\hat{q}_k(x_k)}\,\mathrm{d}x_j\right| = \mathrm{o}_p(n^{-2/5}). \tag{A.12}$$

The approximation (A.12), together with the expressions at (A.6) and (A.11), gives (A.9). Since $\|\hat{Q}\| < \gamma$ with probability tending to one for some $0 < \gamma < 1$, we have

$$\left\|\sum_{s=1}^{\infty}\hat{Q}^s\hat{\mathbf{r}}^A\right\|_{\mathbf{M}} \le \sum_{s=2}^{\infty}\gamma^s\|\hat{Q}\hat{\mathbf{r}}^A\|_{\mathbf{M}} = \mathrm{o}_p(n^{-2/5}).$$

This completes the proof of (A.4). □

Proof of (A.5). Let $\mathbf{l}_1(\mathbf{x},\mathbf{u}) = ((u_1-x_1)m_1'(x_1),\dots,(u_d-x_d)m_d'(x_d))^\top$ and $\mathbf{l}_2(\mathbf{x},\mathbf{u}) = ((u_1-x_1)^2m_1''(x_1)/2,\dots,(u_d-x_d)^2m_d''(x_d)/2)^\top$. To get an idea of which terms in an expansion of $\tilde{\mathbf{m}}^B(\mathbf{x})$ lead to the main terms in the expansion (A.5), we note from an expansion of $m(\mathbf{X}^i)$ that $\tilde{\mathbf{m}}^B(\mathbf{x})$ is approximated by

$$\begin{aligned} &\mathbf{m}(\mathbf{x}) + \hat{\mathbf{M}}(\mathbf{x})^{-1}n^{-1}\sum_{i=1}^{n}\mathbf{Z}^i\mathbf{Z}^{i\top}\mathbf{l}_1(\mathbf{x},\mathbf{X}^i)K_{\mathbf{h}}(\mathbf{x},\mathbf{X}^i)\\ &\quad + \hat{\mathbf{M}}(\mathbf{x})^{-1}n^{-1}\sum_{i=1}^{n}\mathbf{Z}^i\mathbf{Z}^{i\top}\mathbf{l}_2(\mathbf{x},\mathbf{X}^i)K_{\mathbf{h}}(\mathbf{x},\mathbf{X}^i).\end{aligned} \tag{A.13}$$

Define $\tilde{\mathbf{m}}^{B,1}(\mathbf{x}) = \hat{\mathbf{M}}(\mathbf{x})^{-1}\int \mathbf{M}(\mathbf{x})\mathbf{l}_1(\mathbf{x},\mathbf{u})K_{\mathbf{h}}(\mathbf{x},\mathbf{u})\,\mathrm{d}\mathbf{u}$. The second term of (A.13) is then approximated by $\tilde{\mathbf{m}}^{B,1}(\mathbf{x}) + \mathbf{M}(\mathbf{x})^{-1}\sum_{k=1}^d[\partial\mathbf{M}_k(\mathbf{x})/\partial x_k]h_k^2m_k'(x_k)\int u^2K(u)\,\mathrm{d}u$. Also, the third term is approximated by $(h_1^2m_1''(x_1)/2,\dots,h_d^2m_d''(x_d)/2)^\top\int u^2K(u)\,\mathrm{d}u$. Define

$$\begin{aligned}\tilde{\mathbf{m}}^{B,2}(\mathbf{x}) &= \left[\mathbf{M}(\mathbf{x})^{-1}\sum_{k=1}^{d}\frac{\partial}{\partial x_k}\mathbf{M}_k(\mathbf{x})h_k^2m_k'(x_k) + \frac{1}{2}(h_1^2m_1''(x_1),\dots,h_d^2m_d''(x_d))^\top\right]\\ &\quad\times\int u^2K(u)\,\mathrm{d}u\end{aligned}$$

and let $\tilde{\mathbf{m}}^{B,3}(\mathbf{x}) = \tilde{\mathbf{m}}^B(\mathbf{x}) - \mathbf{m}(\mathbf{x}) - \tilde{\mathbf{m}}^{B,1}(\mathbf{x}) - \tilde{\mathbf{m}}^{B,2}(\mathbf{x})$.

For $\ell = 1,2,3$, define $\hat{\mathbf{m}}^{B,\ell}$ to be the solution of the backfitting equation at (2.9) with $\tilde{\mathbf{m}}$ being replaced by $\tilde{\mathbf{m}}^{B,\ell}$. By arguing as in the proof of (A.4), we can deduce that $\hat{m}_j^{B,3}(x_j) = \mathrm{o}_p(n^{-2/5})$ for all $x_j \in (0,1)$. The projection of $\tilde{\mathbf{m}}^{B,2}$ onto $\mathcal{H}(\hat{\mathbf{M}})$ is well approximated by the projection onto $\mathcal{H}(\mathbf{M})$ with a remainder $\boldsymbol{\delta}$ such that $\boldsymbol{\delta}(\mathbf{x}) = \mathrm{o}_p(n^{-2/5})$ for all $\mathbf{x} \in (0,1)^d$. This proves that $\hat{\mathbf{m}}^{B,2}(\mathbf{x}) = \boldsymbol{\beta}(\mathbf{x})n^{-2/5} + \mathrm{o}_p(n^{-2/5})$ for all $\mathbf{x} \in (0,1)^d$.

It thus remains to prove that $\hat{\mathbf{m}}^{B,1}(\mathbf{x}) = \mathrm{o}_p(n^{-2/5})$ for all $\mathbf{x} \in (0,1)^d$. For this bound, we will show that $\hat{m}_j^{B,1}(x_j) = \mu_j(x_j) + \mathrm{o}_p(n^{-2/5})$, uniformly for all $x_j \in [0,1]$, $1 \le j \le d$, where $\mu_j(x_j) = a_j(x_j)/\int K_{h_j}(x_j,u_j)\,\mathrm{d}u_j$ and $a_j(x_j) = m_j'(x_j)\int (u_j - x_j)K_{h_j}(x_j,u_j)\,\mathrm{d}u_j$. For a proof of this claim, it suffices to show that

$$\int \hat{\mathbf{M}}_j(\mathbf{x})^\top [\tilde{\mathbf{m}}^{B,1}(\mathbf{x}) - \boldsymbol{\mu}(\mathbf{x})]\,\mathrm{d}\mathbf{x}_{-j} = \mathrm{o}_p(n^{-2/5}), \tag{A.14}$$

uniformly for all $x_j \in [0,1]$, $1 \le j \le d$. Here, $\boldsymbol{\mu}(\mathbf{x}) = (\mu_1(x_1), \ldots, \mu_d(x_d))^\top$.

We prove (A.14). Note that, uniformly for $x_j \in [0,1]$,

$$\begin{aligned}
&\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \boldsymbol{\mu}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} \\
&\quad = \left[\int q_j(u_j)K_{h_j}(x_j,u_j)\,\mathrm{d}u_j\right]\mu_j(x_j) \\
&\qquad + \sum_{k=1,\ne j} \int \mu_k(x_k)\left[\int q_{jk}(u_j,u_k)K_{h_j}(x_j,u_j)K_{h_k}(x_k,u_k)\,\mathrm{d}u_j\,\mathrm{d}u_k\right]\mathrm{d}x_k \\
&\qquad + \mathrm{o}_p(n^{-2/5}) \\
&\quad = q_j(x_j)a_j(x_j) + \sum_{k=1,\ne j} \int a_k(x_k)q_{jk}(x_j,x_k)\,\mathrm{d}x_k \int K_{h_j}(x_j,u_j)\,\mathrm{d}u_j + \mathrm{o}_p(n^{-2/5}).
\end{aligned}$$

Claim (A.14) now follows from the fact that

$$\begin{aligned}
&\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \tilde{\mathbf{m}}^{B,1}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} \\
&\quad = q_j(x_j)a_j(x_j) + \sum_{k=1,\ne j} \int a_k(x_k)q_{jk}(x_j,x_k)\,\mathrm{d}x_k \int K_{h_j}(x_j,u_j)\,\mathrm{d}u_j + \mathrm{o}_p(n^{-2/5})
\end{aligned}$$

uniformly for $x_j \in [0,1]$. □

A.3. Proofs of Theorems 3 and 4

Recall the definitions of $\hat{\mathbf{M}}$ and $\mathbf{M}$ at (4.2) and (4.8), respectively, in the case of local polynomial fitting. Let $\mathcal{H}_j(\hat{\mathbf{M}})$ denote the space of $(\pi+1)d$-vectors of functions $\mathbf{f} = (f_{j,k})$ in $L_2(\hat{\mathbf{M}})$ such that $f_{j,\ell}(\mathbf{x}) = g_{j,\ell}(x_j)$, $0 \le \ell \le \pi$, for some functions $g_{j,\ell}:\mathbb{R} \to \mathbb{R}$ and $\mathbf{f}_k \equiv (f_{k,0}, \ldots, f_{k,\pi})^\top = \mathbf{0}$ for $k \ne j$. As in the case of local constant fitting, we can write

$\mathcal{H}(\hat{\mathbf{M}}) = \mathcal{H}_1(\hat{\mathbf{M}}) + \cdots + \mathcal{H}_d(\hat{\mathbf{M}})$. Define $\mathcal{H}_j(\mathbf{M})$ likewise. The vectors of functions that take the roles of q_j and q_{jk}, respectively, are

$$\begin{aligned}\boldsymbol{\Psi}_j(x_j) &= \mathbf{N}_1 E(Z_j^2|X_j = x_j)p_j(x_j),\\ \boldsymbol{\Psi}_{jk}(x_j, x_k) &= \boldsymbol{\mu}\boldsymbol{\mu}^\top E(Z_j Z_k|X_j = x_j, X_k = x_k)p_{jk}(x_j, x_k), \qquad k \neq j.\end{aligned}$$

We then have projection formulas analogous to (2.6)–(2.8). For example, for $\mathbf{f} \in L_2(\hat{\mathbf{M}})$ and $\mathbf{g} \in L_2(\mathbf{M})$, we obtain

$$\begin{aligned}(\hat{\Pi}_j \mathbf{f})_j &= \hat{\boldsymbol{\Psi}}_j(x_j)^{-1} \int \hat{\mathbf{M}}_j(\mathbf{x})^\top \mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j},\\ (\Pi_j \mathbf{g})_j &= \boldsymbol{\Psi}_j(x_j)^{-1} \int \mathbf{M}_j(\mathbf{x})^\top \mathbf{g}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j}\end{aligned}$$

and $(\hat{\Pi}_j \mathbf{f})_k = \mathbf{0} = (\Pi_j \mathbf{g})_k$ for $k \neq j$, where $(\hat{\Pi}_j \mathbf{f})_k$ and $(\Pi_j \mathbf{g})_k$ denote the kth $(\pi+1)$-vector of the projection of $\mathbf{f}$ onto $\mathcal{H}_j(\hat{\mathbf{M}})$ and of $\mathbf{g}$ onto $\mathcal{H}_j(\mathbf{M})$, respectively. We can proceed as in the proof of Theorem 1 to prove Theorem 3.

We prove Theorem 4. Decompose $\tilde{\mathbf{m}}$ at (4.3) as $\tilde{\mathbf{m}}^A + \tilde{\mathbf{m}}^B$, where

$$\tilde{\mathbf{m}}^A(\mathbf{x}) = \hat{\mathbf{M}}(\mathbf{x})^{-1} n^{-1} \sum_{i=1}^n \mathbf{v}(\mathbf{X}^i, \mathbf{Z}^i; \mathbf{x})[Y^i - m(\mathbf{X}^i, \mathbf{Z}^i)] K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}^i).$$

Define $\hat{\mathbf{m}}^A$ and $\hat{\mathbf{m}}^B$ from $\tilde{\mathbf{m}}^A$ and $\tilde{\mathbf{m}}^B$, respectively, to be the solutions of the backfitting equation (4.6). It follows that $(\hat{\Pi}_j \tilde{\mathbf{m}}^A)_j(x_j) = \tilde{\mathbf{m}}_j^A(x_j)$, where

$$\tilde{\mathbf{m}}_j^A(x_j) = \hat{\boldsymbol{\Psi}}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n \mathbf{w}_j(x_j, X_j^i) K_{h_j}(x_j, X_j^i) Z_j^i [Y^i - m(\mathbf{X}^i, \mathbf{Z}^i)].$$

As in the proof of Theorem 2, we can prove that $\hat{\mathbf{m}}_j^A(x_j) = \tilde{\mathbf{m}}_j^A(x_j) + \mathrm{o}_p(n^{-(\pi+1)/(2\pi+3)})$ for all $\mathbf{x} \in (0,1)^d$. The stochastic term $\tilde{\mathbf{m}}_j^A(x_j)$ has mean zero and is asymptotically normal. Since $\hat{\boldsymbol{\Psi}}_j(x_j) = \boldsymbol{\Psi}_j(x_j) + \mathrm{o}_p(1)$ and

$$\begin{aligned}&n^{-1} h_j \sum_{i=1}^n \mathrm{var}[\mathbf{w}_j(x_j, X_j^i) K_{h_j}(x_j, X_j^i) Z_j^i Y^i | \mathbf{X}^i, \mathbf{Z}^i]\\ &\quad = \mathbf{N}_2 E[Z_j^2 \sigma^2(\mathbf{X}, \mathbf{Z})|X_j = x_j] p_j(x_j) + \mathrm{o}_p(1),\end{aligned}$$

we find that the asymptotic variance of $\tilde{\mathbf{m}}_j^A(x_j)$ equals

$$n^{-1} h_j^{-1} (\mathbf{N}_1^{-1} \mathbf{N}_2 \mathbf{N}_1^{-1}) \frac{E[Z_j^2 \sigma^2(\mathbf{X}, \mathbf{Z})|X_j = x_j]}{p_j(x_j)[E(Z_j^2|X_j = x_j)]^2}.$$

Next, we approximate $\hat{\mathbf{m}}^B(\mathbf{x})$. Define

$$\tilde{\mathbf{m}}^{B,1}(\mathbf{x}) = \frac{1}{(\pi+1)!}\mathbf{M}(\mathbf{x})^{-1}n^{-1}\sum_{i=1}^{n}\mathbf{v}(\mathbf{X}^i,\mathbf{Z}^i;\mathbf{x}) \\ \times\left[\sum_{j=1}^{d} Z_j^i\left(\frac{X_j^i - x_j}{h}\right)^{\pi+1} m_j^{(\pi+1)}(x_j)h_j^{\pi+1}\right]K_{\mathbf{h}}(\mathbf{x},\mathbf{X}^i)$$

and $\tilde{\mathbf{m}}^{B,2}(\mathbf{x}) = \tilde{\mathbf{m}}^B(\mathbf{x}) - \mathbf{m}(\mathbf{x}) - \tilde{\mathbf{m}}^{B,1}(\mathbf{x})$. As in the proof of Theorem 2, we can show that $\hat{\mathbf{m}}_j^{B,2}(x_j) = \mathrm{o}_p(n^{-(\pi+1)/(2\pi+3)})$ for all $x_j \in (0,1)$. We compute $\hat{\mathbf{m}}^{B,1}(\mathbf{x})$. We can prove that, for all $x_j \in (0,1)$,

$$\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \tilde{\mathbf{m}}^{B,1}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} \\ = \frac{1}{(\pi+1)!}\left[\boldsymbol{\mu}\mu_{\pi+1}\sum_{k=1,\neq j}\int q_{jk}(x_j,x_k)h_k^{\pi+1}m_k^{(\pi+1)}(x_k)\,\mathrm{d}x_k \right. \\ \left. + h_j^{\pi+1}\boldsymbol{\gamma} q_j(x_j)m_j^{(\pi+1)}(x_j)\right] + \mathrm{o}_p(n^{-(\pi+1)/(2\pi+3)}), \tag{A.15}$$

where q_j and q_{jk} are as defined at (A.1) and (A.2), respectively, and $\mu_{\pi+1} = \mu_{\pi+1}(K)$. We also have

$$\int \hat{\mathbf{M}}_j(\mathbf{x})^\top \hat{\mathbf{m}}^{B,1}(\mathbf{x})\,\mathrm{d}\mathbf{x}_{-j} = \boldsymbol{\mu}\boldsymbol{\mu}^\top \sum_{k=1,\neq j}\int q_{jk}(x_j,x_k)\hat{\mathbf{m}}_k^{B,1}(x_k)\,\mathrm{d}x_k \\ + \mathbf{N}_1 q_j(x_j)\hat{\mathbf{m}}_j^{B,1}(x_j) + \mathrm{o}_p(n^{-(\pi+1)/(2\pi+3)}) \tag{A.16}$$

for all $x_j \in (0,1)$. Now, we observe that $\boldsymbol{\mu}^\top\mathbf{N}_1^{-1} = (1,0,\ldots,0)$ since $\boldsymbol{\mu}$ is the first column of $\mathbf{N}_1$. Thus,

$$\boldsymbol{\mu}\boldsymbol{\mu}^\top\mathbf{N}_1^{-1}\boldsymbol{\gamma} = \boldsymbol{\mu}(1,0,\ldots,0)\boldsymbol{\gamma} = \boldsymbol{\mu}\mu_{\pi+1}.$$

Comparing the two systems of equations (A.15) and (A.16), and by the uniqueness of $\hat{\mathbf{m}}^{B,1}$, we conclude that

$$\hat{\mathbf{m}}_j^{B,1}(x_j) = (\mathbf{N}_1^{-1}\boldsymbol{\gamma})h_j^{\pi+1}m_j^{(\pi+1)}(x_j)/(\pi+1)! + \mathrm{o}_p(n^{-(\pi+1)/(2\pi+3)})$$

for all $x_j \in (0,1)$, $1 \le j \le d$. This completes the proof of Theorem 4.

Acknowledgements

Y.K. Lee was supported by National Research Foundation of Korea (NRF) Grant NRF-2010-616-C00008. E. Mammen was supported by the DFG-NRF Cooperative Program. B.U. Park was supported by NRF Grant No. 20100017437, funded by the Korea government (MEST).

References

[1] Bickel, P., Klaassen, A., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: The Johns Hopkins Univ. Press. MR1245941

[2] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95** 888–902. MR1804446

[3] Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear times series. *J. Amer. Statist. Assoc.* **95** 941–956. MR1804449

[4] Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88** 298–308. MR1212492

[5] Connor, G., Linton, O. and Hagmann, M. (2007). Efficient estimation of a semiparametric characteristic-based factor model of security returns. FMG Discussion Paper, Financial Markets Group.

[6] Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26** 943–971. MR1635422

[7] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

[8] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

[9] Fengler, M., Härdle, W. and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *J. Financ. Econ.* **5** 189–218.

[10] Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* **5** 81–102.

[11] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall. MR1082147

[12] Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

[13] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

[14] Huang, J.Z., Wu, C.O. and Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika* **89** 112–128. MR1888349

[15] Huang, J.Z., Wu, C.O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

[16] Lee, Y.K., Mammen, E. and Park, B.U. (2010). Backfitting and smooth backfitting for additive quantile models. *Ann. Statist.* **38** 2857–2883.

[17] Mammen, E., Linton, O. and Nielsen, J.P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496

[18] Mammen, E. and Park, B.U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33** 1260–1294. MR2195635

[19] Noh, H.S. and Park, B.U. (2010). Sparse varying coefficient models for longitudinal data. *Statist. Sinica* **20** 1183–1202. MR2730179

[20] Park, B.U., Hwang, J.H. and Park, M.S. (2010). Testing in nonparametric varying coefficient models. *Statist. Sinica*. To appear.

[21] Park, B.U., Mammen, E., Härdle, W. and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. *J. Amer. Statist. Assoc.* **104** 284–298. MR2504378

[22] Wang, J. and Yang, L. (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Ann. Inst. Statist. Math.* **61** 663–690. MR2529970

[23] Wang, L., Li, H. and Huang, J.Z. (2008). Variable selection for nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

[24] Yang, L., Park, B.U., Xue, L. and Härdle, W. (2006). Estimation and testing for varying coefficients in additive models with marginal integration. *J. Amer. Statist. Assoc.* **101** 1212–1227. MR2328308

[25] Yu, K., Park, B.U. and Mammen, E. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

Received May 2010 and revised September 2010